ESI指标原理及计算

王颖鑫
中国科学院文献情报中心 北京 100080
黄德龙
中国科学院管理决策与信息系统重点实验室 北京 100080
刘德洪
中国科学院武汉文献情报中心 武汉 430071

〔摘要〕针对美国基本科学指标数据库(Essential Science Indicators)中计量指标的设置，在对ESI及其结构进行简单介绍的基础上，分析其指标内涵、设置原理及计算方法；同时介绍引文阈值和指标值的校正，以期正确理解使用ESI所设置的指标，促进对该基本分析评价工具的运用。
〔关键词〕ESI 计量指标 科研评价
〔分类号〕G354.2

Statistical Indicators' Principle and Data Process of ESI

Wang Yingxin
Library of Chinese Academy of Sciences, Beijing 100080
Huang Delong
Key Laboratory of Management Decision and Information System, Beijing 100080
Liu Dehong
Wuhan Library of Chinese Academy of Sciences, Wuhan 430071

〔**Abstract**〕Based on the brief introduction to ESI and its structure, this paper introduces the indicators and its meanings of ESI in details, then it describes the principles for indicators installation and the calculating method. It also shows how to revise the indicators so as to make a full use of the evaluation tool.
〔**Keywords**〕ESI statistical indicators scientific performance evaluation

1 ESI简介

ESI(Essential Science Indicators)即基本科学指标数据库，是由美国科技信息所(ISI)推出的基于SCI和SSCI所收录的全球11 000多种学术期刊的1 000多万条文献记录而建立的计量分析数据库。它是衡量科学研究绩效、跟踪科学发展趋势的基本分析评价工具。进入ESI前50%国家／地区发表的论文占到全球该领域论文的90%以上，用户可以从该数据库中了解到一定排名范围内科学家、科研机构、国家和学术期刊在某一学科领域的论文量和影响力，确定关键的科学发现，评估研究绩效，掌握科学发展的趋势和动向。

2 ESI结构

ESI由引文排位(Citation Rankings)、高被引论文(Most Cited Papers)、引文分析(Citation Analysis)和评论报道(Commentary)4部分构成。

引文排位部分依据论文总被引量，排出位居国际前1%科学家、前1%科研机构、前50%国家／地区以及前50%期刊。高被引论文部分依据论文总被引频次，排出国际1%顶尖论文和1%热门论文。

引文分析部分中Baselines(基准线)用来测度论文组的累积被引频次，Averages(平均被引频次)基于从论文出版年到当前的被引累积数，列出了近10年的年平均被引频次和10年的总平均被引值；以Percentiles(百分点)为被引基准，给出了每一年每一学科领域进入前0.01%、0.10%、1.00%、10.00%、20.00%和50.00%所需达到的引文量；Field Rankings依据总被引频次

收稿日期:2006 02 27 本文起止页码:73-75, 35

给出了22个学科的总体排序，并用图表描述了以5年为单位的连续时间段内该学科的论文数、引文数、篇均被引数变化的数量趋势。

引文分析部分中Research Fronts(研究前沿)基于5年时间段多学科范围内被引频次高的论文，通过聚类分析、共引分析列出各学科领域研究前沿。该部分指标包括高被引论文数(即分析的文献量)、引文数(即高被引论文被引次数，反映了研究前沿的规模)、平均引文数(表明前沿研究的集中度)和平均年份(Mean Year，越近表示研究越前沿)。

评论报道部分对ESI中涉及的特定领域、科研成果等进行采访报道与评论。其中，In-Cites对ESI中重要论文、发现的幕后细节进行报道、评述分析及展望。Special Topics对选定专题领域的文献进行深入剖析并提供相关内容。Science Watch基于热点，追踪基础科学研究领域的发展趋势和研究现状。

3 ESI指标、原理、阈值及计算

ESI处理的数据仅限于ISI收录的期刊论文(科技论文、评述论文、会议论文以及研究报告)。编辑信件、更正通知、摘要、图书、图书的章节以及未被ISI索引的期刊论文，均不被考虑在内。数据一年更新6次，更新周期为2个月。

ESI以引文分析为基础，出版和引文活动可以衡量各国科研水平、期刊的声誉和影响力，也可以反映科研机构和科学家的学术水平。其中，被引频次作为同行认知(Peer Recognition)的一种形式，反映科研群体对科学家的依赖程度。

3.1 论文数与引文数

论文数是描述科学家、期刊、机构、国家发表论文能力的一个基本指标，是在给定时期或给定领域内发表或刊载论文的数量。引文数是从使用者的角度评价科学家、期刊、机构、国家科学水平的一个基本指标，是论文被引用的全部次数，它用客观使用的数量反映了科学体在科学发展和文献交流中的作用。这两个指标都是绝对数量指标，一般来说，其值越大，表明该科学家、期刊、机构、国家的作用越重要。

ESI将论文数和引文数作为指标，针对不同对象，对期刊论文的第一作者和非第一作者平等对待，论文引用和被引频次平等归于所有作者，体现了对科学参与者的公平评价。时间段为10年(包括当前更新时间)，从ISI收录该论文的实际年份算起，反映了文献从发表到引用高峰再到引用稀少的客观过程。热点论文计算的时间段为2年，计算国际上过去2年中各领域论文在近2个月被引用的次数，也是依热点问题的生命周期和人们的关注程度而定。

3.2 篇均被引频次和平均被引频次

篇均被引频次(Average Citations Per Paper)是给定时间内，期刊所载文献被引数量除以该刊全部论文数。以科学家为例，它表示科学家所发表每篇论文被引用的平均水平，其值高则一般代表该科学家水平高；它同样适用于用于期刊、机构和国家。作为一个相对数量指标，它弥补了绝对数量指标中马太效应导致的偏差。在ESI中篇均被引频次即引文数除以论文数，表示每篇论文被引用的平均水平，针对不同对象，篇均被引频次反映该对象的学术水平高低。

ESI中平均被引频次(Averages)与篇均被引频次不同，ESI的Baselines中给出了各领域论文每年的年平均被引频次和10年累积平均被引频次，Averages值由某领域总引文数除以总论文数得到。这些平均值可以被用作科学家、机构、国家以及期刊排位表给出的单篇被引值的基线，独立年份的学科领域平均值可用于该年份出版的论文的比较。

3.3 平均年份(Mean Year)

该指标出现在Research Fronts中，它是引文发表的平均年份，是衡量学术界对相关主题研究的活跃(Currency)程度的一个指标，其核心思想是：引文发表的平均年份越近，表示当前对该主题开展的研究越多。Mean Year离当前年份越近越能表明该主题处于当前学科热点或研究前沿。即Mean Year就是前沿课题研究兴起的时间点。

计算其值，要从论文的发表年份开始到当前引用年份，将年月转变成数字：1－12月分别对应0，1/12，2/12，……11/12，年份为整数部分，然后对所有数字求算术平均即可。

3.4 标准共引阈值(Normalized Co-citation)

引用表现学科领域间的联系，共引反映科学领域内重要问题之间的联系。所谓共引，是在给定论文的参考文献中，对某一论文的引用伴随着对另一论文的引用。Research Fronts是引文网络结构根据若干篇原创性成果的核心文献来描述某个特定研究领域现状的应用。它汇集特定领域核心文献和研究焦点，追踪学科发展趋势，辨析科学家、研究机构、国家对科学发展的贡献。

ESI采用单连接聚类算法(Single-Linkage)，其基本思想是两个簇之间的距离为从两个簇中抽取的每对样本的最小距离。通俗地讲，样本点离哪个类近就划入哪一类，表达关系密切、性质相近的意思。Research Fronts的聚类分析以共引强度为基本计量单位，分析之前需要先为论文设定共引强度阈值(Integer Co-citation Frequency)，目的是去除大量弱相关论文(噪音)，然后形成学科强相关的论文簇，进而定量分析。为筛选具有一定共引强度的论文，设定了标准共引阈值。假设有论文A和B，其共引阈值的计算公式如下[11]：

标准共引阈值＝论文A和B的共引强度阈值／(论文A的引文数×论文B的引文数)$^{0.5}$

该公式通过聚类分析推导而来，其中共引强度阈值一般以专家打分的方式给出，然后通过该公式转换成标准共引阈值。ESI在处理数据时将Integer Co-citation Frequency赋值为2，为Normalized Co-citation赋予0.3的值。

3.5 引文阈值

引文阈值作为筛选标准，用来从各领域中选出一定比例的科学家、科研机构、国家和期刊，引文数大于等于阈值者均可入选。针对不同学科、学科特点及引文率的不同，各领域设定不同的引文阈值。ESI设定了国际顶尖论文引文阈值，考虑到学科不同和时间上新旧文献的可比性，将每个学科每年分别设定不同的值，将某论文10年内累积引文数与阈值比较，大于等于则可以入选。热点论文引文阈值，每个领域每两个月设定不同的值，将某论文2个月内的累积引文数与阈值比较，大于等于则可以入选。阈值如表1、表2、表3，表中所有数据均为2005年9月1日更新：[2]

表1 ESI引文阈值(1995年1月-2005年6月)

引文阈值 比较项 / 学科领域	科学家	国家	研究机构	期刊
农业科学	173	178	646	724
生物与生物化学	800	243	4 411	1 932
化学	710	481	2 944	2 047
临床医学	1 159	1 643	1 549	2 291
计算机科学	95	41	578	345

表2 ESI国际顶尖论文引文阈值(1995年1月-2005年6月)

引文阈值 年份 / 学科领域	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
农业科学	64	57	55	55	50	43	32	22	15	6	3
生物与生物化学	218	206	198	170	146	120	93	69	44	18	4
化学	106	101	92	86	79	71	54	44	26	13	3
临床医学	173	152	141	128	113	98	78	59	36	15	4
计算机科学	51	48	43	44	37	29	26	22	11	5	

表3 ESI热点论文引文阈值(2003年8月-2005年6月)

引文阈值 年份 / 学科领域	2003-4	2003-5	2003-6	2004-1	2004-2	2004-3	2004-4	2004-5	2004-6	2005-1	2005-2	2005-3
农业科学	8	8	9	7	9	4	6	6	5	4	3	5
生物与生物化学	18	18	20	17	18	17	15	15	9	7	5	5
化学	13	14	14	13	14	12	11	11	9	7	5	5
临床医学	20	16	17	17	16	21	16	12	10	8	6	5
计算机科学	7	11	7	6	8	11	6	5	4	4	3	6

* 热点论文阈值每2个月更新一次，表中"-4"、"-5"…"-3"表示次数。

3.6 All Fields(全部领域)

ESI多处从"All Fields"全学科领域的角度出发，向研究者展示科学全貌、科学前沿全貌。引文排位部分里，直接在"Select a Scientist From This Field"的下拉列表中选择"All Fields"点击"GO"按钮，不分领域依据引文数列出了居前科学家，由此可以观察判断排在科学发展前列的科学家及其所研究的领域，也可以判断学科研究间的关系。同理，对于科研机构、国家/地区以及期刊，高被引论文部分"All Fields"可以判断出科学发展的热点和重点。

3.7 跨学科期刊归类

ESI中任何一种专业期刊都只能归入唯一领域。对于跨学科期刊，如*Nature*、*Science*等，之前全部归入Multidisciplinary Field。现在ESI对约60种跨学科期刊论文按照其引文对其进行归类，即论文的归类取决于其引文和参考文献的归类。其主要依据引文的特点，引文是一篇论文对另一篇论文的应用，一定程度上反映学科内容之间的引证关系。例如，一篇刊载在跨学科期刊上的论文，如果其大多数引文属于神经系统科学(Neuroscience)领域，且大多数参考文献来自神经系统科学领域，那么该论文就被归入Neuroscience。

跨学科期刊论文的归类情况因期刊而异，如*Nature*、*Science*的再分类率可达95%。采取此法，60种期刊约17万篇论文中近半数被归入具体领域。重新归类为科学家、机构、国家、期刊论文排序提供了更准确的统计数据，可更准确地反映各学科领域的研究情况、学术成果、影响力。

3.8 ESI指标值校正

ESI对科学家、科研机构、国家和期刊在一定时期内分别进行排序。时间序列以5年为一段，有部分重叠依次连续后推，即1995-1999年、1996-2000年。这样采用5年期的移动平均(Moving Average, MA)方法对一个科学家、机构、国家或期刊的科研能力进行评估，旨在减少异常值的影响。这样评价保持了评级结果的稳定性。

ESI每2个月更新一次，所以当前年份各项指标值都不够完全，不能反映实际情况。ESI根据长期观察总结发现：假设有稳定的出版量，整个数据库中，每5年最后一年的引文数平均占该时期全部引文数的41%，而这41%的引文又近似平均分布在这一年的6个时间段内。据此估算得到表4：[3]

表4 ESI参数

双月周期	1	2	3	4	5	6
引文数	1.52	1.37	1.26	1.16	1.07	1.00
论文数	1.20	1.15	1.11	1.07	1.03	1.00
影响因子	1.26	1.19	1.13	1.08	1.04	1.00

根据此参数表来校正，例如ESI今年第三次更新，且引文数、论文数和篇均被引频次分别是20、10、2.0，则全年引文数、论文数和篇均被引频次分别是$20 \times 1.26 = 25.2$、$10 \times 1.11 = 11.1$、$2.0 \times 1.13 = 2.26$。第六次校正参数为1，表明在该年最后已经没有估算必要。

4 结 语

ESI这一基本分析评价工具的核心即是其指标，这些指标的设置基于统计、文献计量等知识，经过长期观察筛选而形成。在应用和研究ESI的过程中，只有明确这些指标的原理、计算和意义，才能得出明晰的结论。

参考文献：

1 Research front methodology.[2005-10-13].http://www.esi-topics.com/RFmethodology.html

2 Citation thresholds.[2005-10-13].http://www.in-cites.com/thresholds.html

3 Graphs used in essential science indicators: Projecting full-year citation counts.[2005-10-13].http://www.in-cites.com/graph.html

SCI收录的期刊上发表文章为荣，有些机构甚至不惜重金奖励这种发表行为，而忽视了这些论文真正的科学价值。试想，爱因斯坦在发表他“相对论”的理论发现时，是否考虑过期刊是否是核心期刊呢？由于一些核心期刊的“泛滥”，很多科学家转而选择国际会议发表自己的论文，而实际上，国际会议也存在“泛滥”的势头。

社会化引文网络不是要在SCI之外新建一个开放的引文数据库，而是致力于建立一种开放、动态和透明的机制。在这种机制和环境下，科学家能够通过引文更加专注论文的科学主题，而不是过分关注科研项目回馈或者由此所得到的社会声望。因特网和信息技术的飞速发展也为自动化引文加工处理机制提供了可能，而这一切将促使科学家们不断完善自己的引文行为。

参考文献：

1 Perkel J M. The future of citation analysis —— The challenge is to track a work's impact when published in nontraditional forms.[2005-11-29].http://www.the-scientist.com/2005/10/24/24/1
2 Open Citation (OPCIT) Project.[2005-11-29].http://www.ecs.soton.ac.uk/~harnad/citation.html
3 Hitchcock S et al. Open citation linking: The way forward. D-Lib Magazine, 2002:8(10).[2006-02-05].http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html
4 Small H. Why authors think their papers are highly cited. Scientometrics, 2004,60(3):305-316
5 Small H. Belver and Henry. Scientometrics, 2001,51(3):489-497
6 Small H. On the shoulders of Robert Merton: Towards a normative theory of citation. Scientometrics, 2004,60(1):71-79
7 Cilibrasi R, Vitanyi P M B. Automatic meaning discovery using Google.[2006-02-05].http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cs/0412098
8 Noruzi A. The web impact factor: A critical review. The Electronic Library.[2006-02-05].http://eprints.rclis.org/archive/00005543/01/Web_Impact_Factors,_A_critical_review.pdf
9 Ball P. Index aims for fair ranking of scientists. Nature, 2005, 436(900):doi:10.1038/436900a.[2006-02-05].http://www.nature.com/nature/journal/v436/n7053/full/436900a.html
10 Kuhn T. S. The structure of scientific revolutions, 2nd ed. Chicago: University of Chicago Press, 1969.174-210
11 Small H. Visualizing science by citation mapping. Journal of the American Society for Information Science, 1999,50(9):799-813
12 Small H. A general framework for creating large-scale maps of science in two or three dimensions: The SciViz system. Scientometrics, 1998,41(1-2):125-133
13 Cronin B. Bibliometrics and beyond: Some thoughts on web-based citation analysis. Journal of Information Science, 2001 (27):1-7

〔作者简介〕 毛 军，男，1974年生，副研究馆员，管理学博士，发表论文17篇。

（上接第75页）

4 霍艳荣. 基本科学指数(Essential Science Indicators)数据库. 图书情报工作, 2003,47(1):56-59
5 刘 清，郎 荣，李军虹等. 美国《基本科学指标》的结构及其应用. 情报杂志, 2004(5):94-96
6 庞景安. 科学计量研究方法论. 北京：科学技术文献出版社, 2002
7 Essential acts.[2005-10-13].http://www.in-cites.com/all-essential-facts-list.html
8 Essential science indicators.[2005-10-13].http://scientific.Thomson.com/products/esi/
9 Classification of papers in multidisciplinary journals.[2005-10-13].http://www.in-cites.com/class-multi-jnl.html
10 Yancey R. Essential science indicators: Thomson Scientific Evaluation Tool delivers more precise measurement.[2005-10-13].http://scientific.thomson.com/press/2005/8288727/
11 Essential science indicators data information.[2005-10-13].http://www.in-cites.com/ESI_Product_Info/

〔作者简介〕 王颖鑫，女，1982年生，硕士研究生，发表论文2篇。
黄德龙，男，1981年生，博士研究生，发表论文6篇。
刘德洪，男，1962年生，研究员，发表论文15篇。