美国《基本科学指标》的结构及其应用

刘　清　　邵　荣　李军虹　李　珑　暴朝霞

（武汉大学信息管理学院）　　（中国科学院武汉文献情报中心）

摘　要　介绍了基本科学指标的主要模块、数据处理规范及数据入选标准，并对其具体应用作了简要分析。

关键词　基本科学指标　美国

基本科学指标(Essential Science Indicators，以下简称ESI)是美国费城科学情报所(Institute for Scientific Information，以下简称ISI)"研究服务组"2001年推出的定量化程度极高的数据库。它是一种基于网络的文献引证分析环境，是定量地评估科学研究水平的重要研究工具。它通过ISI Web of Knowledge提供服务，是ISI网络集成服务平台的一个重要组成部分。

1　ESI的结构

1.1　ESI的主要模块　ESI的内容均派生于ISI的数据库，包括引文排位(Citation Rankings)、高被引论文(Most Cited Papers)以及引文分析(Citation Analysis)三大主要模块。其中引文排位模块包括高被引(most cited)作者、机构(大学、公司、政府研究机构)、国家和期刊排位表，高被引论文模块包括高被引论文和热点论文列表，引文分析模块包括基线(baseline)和研究前沿列表。

除了上述的3个主要模块之外，ESI还提供其它内容，包括：通过对数据分析的导引与诠释的编辑简评来强化ESI提供的各种表格和数据集；对被选的研究领域给予了特别关注、被称为"特殊话题"的编辑点评。下文对这些模块分别予以介绍。

1.1.1　科学家排位。被引频次是同行认知(peer recognition)的一种形式，通常反映的是科研群体对于科学家的依赖程度。甚至可以说，高被引科学家形成科研群体的实质核心。许多高被引科学家以荣誉奖项的方式获得同行认知。ESI根据科学家论文被引频次的总和对科学家进行排位。依据总被引频次，排位科学家属于前1%范围以内。

1.1.2　机构排位。科学研究是在科研机构之内进行，因而对于与研究机构有关系的科研人员的认知就反映在研究机构的整体声誉当中。通过合计研究机构层级的出版物与引文频次，可以衡量机构产出与机构声誉。ESI根据各机构论文的被引频次的总和对机构进行排位。依据总被引频次，排位机构属于前1%范围以内。

1.1.3　国家排位。如同出版与引文活动所衡量的，各国科研成就的水平参差不齐，科研成就的分布是不平衡的。一个国家的科研活动水平，大致上与该国的国民生产总值(GNP)或者其它经济产出的能力有关联。进行单篇论文被引频次的国家比较有助于纠正国家规模与论文产出上的差异。ESI根据各国论文的被引频次的总和对国家进行排位。依据总被引频次，排位国家属于前50%范围以内。

1.1.4　期刊排位。同样地，期刊在声誉及影响方面存在差异，这反映在期刊的被引频次上。ESI提供了长期的期刊引文排位。我们还可以通过查询ISI出版的《期刊引证报告》(Journal Citation Reports，JCR)与短期引文行为进行比较。ESI根据期刊发表论文的被引频次的总和对期刊进行排位。依据总被引频次，排位期刊属于前50%范围以内。

1.1.5　高被引论文(highly cited papers)。ESI根据论文的被引频次，选择靠前的1%范围内的论文形成高被引论文列表。一般地，论文的被引频次的高峰出现在论文发表后的第2～4年，某些论文则被持续引用多年。少数文献有着延迟的认知。模式上差异很大，这与论文的形式、所属领域以及所报道的发现的性质有关。例如，报道发现的论文的被引频次上升很快，随着其它文献对发现的进一步阐述，其被引频次也会很快下降。而报道方法与技术的文献的被引频次则随着方法与技术的传播以及用途的证实逐渐上升。ESI设定了相对于特定领域与年份的不同的被引频次标准，保证入选的论文在相应的领域和年份里，其被引频次属于靠前的1%范围以内。

1.1.6　热点论文。热点论文指的是与相同领域与出版年的其它论文相比，出版后很快就有高被引频次的论文。热点论文的选择也是基于一定的条件。但是选择的时间段相当短，亦即，论文的出版年龄不能超过2年，而且是在当前的2个月里被引。这意味着论文必须在很近的一段时间里得到关注。每一领域及时间段都设定了入选条件，按照相应的条件，0.1%的论文得以入选。

1.1.7　基线(BASELINES)。基线是横跨大的论文组的累积被引频次的测度，这些大的论文组具有一定被引频次。基线由"平均被引频次"和"百分点"两个表格组成。

a.平均被引频次。ESI计算1992～2002年共10年里每一年的平均被引频次，这基于从论文出版年到当前的被引累计数。平均值等于单篇论文被引频次之和除以论文总数。全部10年的平均数在"all years"栏中给出。各单独学科领域以及全部学科领域的被引频次均被给出。物理学1991年10.3的平均被引频次，含义是，物理学领域1991年出版的论文从该年到现在平均每篇被引用10.3次。各领域(或者全部领域)的10年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以用于进行该年份出版的论文的比较，无论这些文献属于ESI列出的高被引论文，还是来源于《科学引文索引》(Science Citation Index，SCI)的论文。

b.百分点(percentile)。"百分点"表示的是一个被引基准，在

这一基准或者高于这一基准，论文的固定比例开始下降。百分比通常是 1%，用在这里表示按照被引频次排序的顶尖论文的固定比例。ESI 选定的学科领域和年份的百分比数字有 0.01%、0.1%、1%和 10%。

被引频次是高度倾斜的，许多论文并非频繁地被引，相对地，高被引论文仅占很少的一部分，这大概是一种铁定的分布规律。选择的方法之一是将论文按照被引频次降序排列，然后选择靠前的一定比例的论文。百分比表显示的是各领域、全部领域和各年份的 4 个不同的百分比标准的被引频次入选条件。例如，1993 年航天物理学论文被引频次为 44 的入选条件，选出的是占 1993 年航天物理学期刊出版的论文总数 1%的论文。

1.1.8 研究前沿。科学研究区域，特别是那些学科发展前沿的区域，特点表现在科学家之间紧密联系的模式上。模式以多种方式展现，包括正式的和非正式的。其中突出的是一个科学家对另一个科学家工作的引用。引用的形式反映了科学家如何在其他人工作的基础上达成自己目标的精心选择的过程。这种引文网络结构可以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的现状。

在 ISI 的基本科学指标中达成这一目标的工作成为"研究前沿分析"(RESEARCH FRONT ANALYSIS)。它是基于确定 5 年时间段多学科范围内的被引频次高的论文，然后确定这些论文被共同引用的程度。即是，在给定论文的脚注或者参考文献中，对某一条目的引用是否伴随着对另一高被引条目的引用。这用于确定两篇高被引论文的同被引频次。

1.1.9 In-Cites。除各种排位以及其它计量数据外，ESI 还为我们提供定期更新的编辑点评材料的选本。这包括 In-Cites，一种收录与 ESI 收录的各数据类型有关的述事与述评的文集。在专为 ESI 所写的以述事和原始述评为特色的访谈栏目里，科学家们讲述有关他们的高被引论文的幕后花絮，以及关于他们所在领域的应用展望以及其它未来发展的述评。此外，"特写"(feature articles)还指出新兴(emerging)的学科领域、高被引研究机构、不同国家的研究状况、具有高影响因子的期刊以及其它的话题。

1.1.10 特殊话题(special topics)。特殊话题基于对选定的专题领域内文献的分析，对该领域给予深度地审视。点击某一给定话题，将会唤出有关该话题的数据，这些数据包括领域的成长、该领域的高被引论文、科学家、研究机构以及国家等方面。特殊话题代表的是一种较窄的文献标准。ESI 采用一种联合词汇检索和研究前沿分析结合的方法来确定话题。

1.1.11 《科学观察》(Science Watch)。ESI 也包含一年前的《科学观察》的材料。《科学观察》是一种双月简报，特点是基于最近和最热的基础研究的引文分析。其典型内容是：选定领域的高被引科学家、机构排序，检视热点领域或者新兴领域，跟踪国家或者国际科研趋势。另一个特色是刊登国际顶尖科学家的访谈录。每一期包括生物学、医学、物理学以及化学的十大热点论文，这些热点论文在出版两年内，在最近的 2 个月里有着相当高的被引频次。每一列表都伴随专家述评。

1.2 ESI 数据处理的规范 ESI 处理的数据仅限于 ISI 索引的期刊论文。图书、图书的章节以及未被 ISI 索引的期刊论文，无论是以出版计数而言还是以引文计数而言，均不被考虑在内。数据每 2 至 4 个月更新一次。

1.2.1 论文及被引频次与作者、机构、国家或期刊的关系。按照 ESI 的解释，根据研究对象的不同，一篇论文平等的归于所有作者，或者作者署名的机构、国家，或者出版该论文的期刊，该论文的被引频次亦平等地归于所有作者、机构、国家或者期刊。在 ESI 的统计过程中，第一作者、第一作者所在的机构与其他作者以及他们所在的机构同等对待，并不会被加上特殊的权重。亦即，ESI 不对第一作者和非第一作者进行区分。

1.2.2 计数的时间段(被引频次、论文、单篇论文被引频次)。ESI 计算被引频次的时间段是 10 年，外加当前年度的实际月份(数据每 2~4 月个更新一次)。这就意味着，任何在这一 10+的时间段里的论文是在相同的时间段里被引用。对于热点论文而言，ESI 用 2 个月而不是 12 个月来计算热点论文的年龄。他们仅仅检视过去两年出版的论文，了解某些论文是否获得了非同寻常的被引频次。

1.2.3 被计数的文献类型。前面已经提到，ESI 只计算 ISI 索引的论文。它将论文定义为常规科研论文、综述论文、会议论文以及研究札记，致编者信、更正以及文摘不被计算在内。

1.2.4 包含的期刊。ESI 计数基于一个分为 22 大领域的 ISI 期刊集。按照期刊的唯一归类来定义这些领域，即任何一种期刊不可能在归入某个领域的同时又归入其它领域。跨学科领域(Multidisciplinary field)包含诸如《科学》、《自然》之类的杂志。按照最新的规则，ESI 对约 60 种跨学科类杂志的具体文献按照它们的引文进行重新分类，其中近半数被归入具体的领域。

2 ESI 数据入选条件

2.1 科学家/国家/研究机构/期刊排位 ESI 科学家、研究机构、国家以及期刊排位基于 10 年期的高被引科学家、研究机构、国家以及期刊的被引频次，只有符合一定条件的对象才可以入选这一排位表中。为了表达学科领域间被引频次的差异，ESI 对不同的学科领域应用不同的入选条件。按照各学科领域的入选条件，大致上各领域有相同比例的文献入选。对于科学家领域而言，这个比例是 1%。按照这一比例，每一个学科领域可以诠释出特定的被引频次标准。研究机构排位表的比例也是 1%。对于国家及期刊排位表，这个比例是 50%。表 1、表 2 给出的是科学家/国家/研究机构/期刊引文数入选条件，以及各研究对象入选比例(由于篇幅所限，下面的表格仅给出部分学科的入选条件)。

表 1 科学家/国家/研究机构/期刊排位入选条件

学科领域	科学家	国家	研究机构	期刊
农业科学	147	111	480	597
生物学与生物化学	832	216	4334	1299
化 学	629	400	2603	1447
计算机科学	77	29	471	242
工 程	171	130	505	372
数 学	130	47	1060	647
物理学	1866	415	3779	1696

表 2 科学家/国家/研究机构/期刊排位入选百分比

研究对象	入选百分比(%)
科学家	1
机 构	1
国 家	50
期 刊	50

2.2 高被引论文 在选择高被引论文的时候，同样也设定了

入选条件。不但有按学科领域的入选条件，而且还有按年份的入选条件。设定各年份的入选条件是为了表达新老文献之间可比性，对于每一领域及时间段，被选择的是前1%的论文。表3显示的是每一领域及相应年份与这一比例对应的被引频次入选条件。

表3 1992年～2002年12月高被引论文入选条件

学科领域	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
农业科学	53	55	51	46	38	33	30	24	16	8	3
生物学与生物化学	227	212	191	173	150	138	108	79	53	26	7
化 学	101	102	91	81	73	61	52	41	28	13	5
计算机科学	48	37	37	36	31	25	21	15	9	6	
工 程	52	46	41	35	32	29	22	17	12	6	
数 学	44	38	37	31	27	22	17	14	9	5	
物理学	117	106	96	87	83	68	57	46	33	17	5

由于被引频次因学科领域的不同而不同，以及较老的文献的被引频次要多于近期的文献，高被引论文的选择首先是计算不同被引频次水平的论文数量，寻求每个领域及年份的分布状况，然后根据分布状况通过选定相同比例的论文来设定选择条件。

特定领域与年份的被引频次标准应用于期刊集的所有论文以选择高被引论文。被引频次条件基于被引频次的分布，它用于选定各领域和年份的特定的顶尖部分的论文。条件基于基线的"百分点"表格的"All Years"栏的数据。

2.3 热点论文 前面在1.6节里已经提到热点论文入选条件，这里不作赘述。简言之，热点论文应当是在最近2年里出版，在当前2个月里有着异乎寻常的高被引频次。

表4 2000年12月～2002年12月热点论文入选条件

学科领域	2001-1	2001-2	2001-3	2001-4	2001-5	2001-6	2002-1	2002-2	2002-3	2002-4	2002-5	2002-6
农业科学	6	5	9	5	5	5	3	3	3	2	2	8
生物学与生物化学	16	18	13	13	11	13	14	9	10	6	5	3
化 学	10	9	9	8	8	8	9	7	5	4	3	2
计算机科学	8	6	8	6	4	4	35	3	2	2		5
工 程	5	5	4	5	4	4	3	3	3	3	3	3
数 学	4	4	4	3	4	3	3	3	3	2	2	2
物理学	10	14	12	12	10	13	8	8	7	6	6	4

3 ESI的应用

虽然ESI诞生的时间很短，但是已经得到比较广泛的认知。这除了ISI自身的推介之外，更主要的还是在于ESI数据自身的特点和优势。它是定量地评估科学研究水平的重要研究工具，可以帮助我们分析各个研究领域中科学发现的影响和趋势，分析研究机构、国家和学术期刊在一定研究领域内的学术影响，辨析重大科学发现，评价科学行为，追踪重要科学进展，并有助于我们对国际科学文献进行系统、客观的分析。

由于ISI只提供数据，不对数据做分析和评价，用户在使用ESI时，就有很大的自由发挥的空间。也就是说，面对相同的数据集，用户可以作出不同角度的诠释、评价和分析，从而达到情报研究、科研评价的目的。在这一点上，ESI具有其它类型信息源所不具备的优点。

国内外已经开始有机构利用它进行科研评价。基于研究近十年来物理学文献引证规律、为科研人员和科研决策人员提供科研与决策依据的目的，中国科学院武汉文献情报中心在ESI的基础上提取了与物理学有关的数据1万多条，结合ESI的重要内容和研究成果，按照文献引证的分析方法，对1992年以来物理学文献的论文数量、引证数量进行统计与分析，形成《1992～2002年物理学发展态势》报告。通过这份报告，可以让科研人员了解在过去的十年里，物理学领域文献发表与引证的特点，从而间接了解不同的国家、期刊、机构乃至科学家在物理学发展过程中作用与贡献，也可从中比较差异、寻求热点。

中国科学院文献情报中心新近完成的《世界科学前沿(2002年)》也是一份基于ESI的研究报告，它为科研工作者提供了世界科学前沿核心论文、国际一流科学期刊以及全球高影响力科学家等方面的内容。

中国科学院资源环境科学信息中心在2002年中连续推出多份基于ESI的数据统计分析报告，它们是：《中国暨中国科学院的科学论文产出与科学影响力调查——以ISI基础科学指标(ESI)的统计分析为例》、《国际地球科学与中国地球科学十年发展态势(1991～2001)——以SCI和ESI的统计分析为例》、《国际生物科学学科发展态势》、《大地测量学国内外机构论文发表情况及学科发展态势》、《国际及中国地球科学发展态势(2002年4月数据分析报告)》、《国际科学十年发展态势(1992～2001)——根据2002年5月ESI数据统计分析》、《国际生物科学十年发展态势(1992～2001)——根据2002年5月ESI数据统计分析》。这些报告分析了科学文献产出的学科分布、国家(地区)分布、年度变化，对中国和国际科学文献产出的学科分布、顶尖论文数量进行了比较。

4 结束语

由于ESI诞生的时间较短，其数据的作用有待进一步地研究，其价值也有待挖掘。笔者想强调的是，由于引文行为自身的特点，在分析被引频次的时候，必须认识到，ESI作为一种基于被引频次的分析工具，有着先天的局限性。《自然》杂志2002年2月的一篇关于引文分析的专栏文章对于将引文分析用于科研评价的作用提出了质疑，认为引文分析"在外行人手里，完全是笨拙的工具"，即便"在专家眼里，原始引文数据也常常包含错误"。此前，《自然》2000年2月的一篇题为"对引文索引的依赖损害了生物多样性的研究"的读者来信，对于引文分析的滥用提出了批评。因此，我们建议，在使用ESI的过程中，不要片面或者局部地看待任何数据，要注意与其它形式的分析与评价手段相结合。只有这样，才能真正地发挥ESI数据的作用。

参考文献

1 ISI Essential Science Indicators. http://isi2.isiknowledge.com
2 Antonio G. Valdecasas, Santiago Castroviejo, Leslie F. Marcus: Reliance on the Citation Index Undermines the Study of Biodiversity. Nature, 2000
3 David Adam: ADAM Citation analysis: The counting house. Nature, 2002
4 中国科学院资源环境科学信息中心：国际生物科学十年发展态势，2002
5 中国科学院资源环境科学信息中心：国际及中国地球科学发展态势，2002
6 中国科学院资源环境科学信息中心：国际科学十年发展态势，2002
7 中国科学院文献情报中心：世界科学前沿(2002年)，2002
8 中国科学院武汉文献情报中心：1992～2002年物理学发展态势，2002

(责编：勃梓钧)