●宋丽萍（天津师范大学信息管理学系　天津　300070）

基于 Web 的学术信息资源引文索引与分析体系

摘　要： 阐述了基于 Web 环境下的学术信息资源引文索引与分析体系——Web of Knowledge，Citeseer 和 Citebase 的功能，并进行了比较。在此基础上，指出了开放环境下引文索引的意义及其发展方向。

关键词： Web；信息资源；引文索引；引文分析/Citeseer；Citebase；SCI

Abstract: This article discusses and compares the functions of Web-based academic information resource citation index and analysis systems, that is , Web of Knowledge, Citeseer and Citebase. Based on this, it points out the significance and future of Web-based citation index in the open access environment.

Keywords: Web; information resources; citation index; citation analysis/Citeseer; Citebase; SCI

著名英国学者 J. M. Ziman 指出：没有一篇科学论文是孤立存在的，它是被深嵌在整个学科文献体系之中的[1]。正是由于文献这种引用关系构成了科学论文网络，并为科学论文检索提供了线索，同时构成了引文分析的基础。Web 中类似于引文的是网页上的超级链接。超级链接的出现使参考文献链接具有了无限扩展能力，只需轻轻点击即可实现一步到位式的参考[2]。1997 年，即 Open Journal 计划之后，美国科技情报研究所（ISI）推出的 Web of Science（WoS）、2004 年改版的 Web of Knowledge，1998 年美国普林斯顿大学 NEC 研究院开发的 Citeseer（又名 Researchindex），以及 1999 年英国南开普顿大学为期 3 年的 Open Citation Project 的产物 Citebase，为以 Web 为基础的引文索引以及引文分析做出了较好的尝试，并为学术论文检索和评估开发了重要的资源。本文阐述了上述三个系统功能，并进行了比较。

1　三个系统的功能

1.1　ISI 的 Web of Knowledge

Web of Knowledge 是 ISI 于 2001 年 5 月推出的学术信息资源体系，2004 年，ISI 对其进行了改版。新版 Web of Knowledge 内容详见表 1。

表 1　Web of Knowledge 的构成

数据库产品	分析工具	其他资源
Web of Science（SCI、SSCI、A&HCI 三大引文数据库，以及 Index Chemicus、Current Chemical Reactions 两大化学数据库） Current Contents Connect（现刊题录快讯数据库） ISI Proceedings（ISTP、ISSHP 两大会议录索引） Derwent Innovation Index（德温特世界专利创新索引） INSPEC（包括物理、工程、电子和计算机科学）	Journal Citation Reports Essential Science Indicators	ISI highlycited. com www. thomsonisi. com

1.1.1　Web of Knoeledge 的检索功能　由于 Web of Knowledge 以 WoS 作为其信息门户，因而本文以 WoS 为例，说明其检索功能。WoS 提供 Quick Search，General Search，Cited Reference Search，Advanced Search 功能。Quick Search 融合了 2001 年版中 Easy Search 的功能；General Search，可以通过主题、著者、来源刊名或著者地址检索来源文献，在 General Search 界面的检索框内输入检索词后，可对检索结果进行语种、文献类型或排序的限定；Cited Reference Search 即被引文献检索式，可通过被引著者、被引著作和被引文献发表年份进行检索，这是 WoS 中最独特的检索途径；Advanced Search 为复杂提问式所采用。WoS 提供了 3 个重要的链接：Times Cited，Cited References 以及 Related Records。其中，Times Cited 提供了该文献在数据库中所有引用文献列表及被引次数；Cited References 提供了该文所有参考文献的记录；Find Related Records 即如果两篇论文共同引用了同一篇或几篇参考文献，则该两篇论文构成相关记录，同时显示共引数量。正是通过这 3 种重要链接，WoS 将整个记录组织成一个网状的结构，有效地整合了不同学科的数据库，并通过“View Full Text”链至本馆已定购的电子出版物全文，通过“Holdings”直接链至本馆 OPAC 系统显示馆藏，形成以知识为基础的学术信息资源整合体系[3]。

1.1.2　Web of Knowledge 引文分析功能　Analytical Tools 由期刊引文报告（Journal Citation Reports），基础科学指标数据库（Essential Science Indicators，ESI）构成，其中 ESI 为 2001 年 ISI 基于 ISI 引文索引而建立的计量分析数据库，

同时也是衡量科研绩效、跟踪科学发展趋势的基本分析评价工具[4]。其中包括引文排序（Citation Rankings），高频被引文献（Most Cited Papers）以及引文分析（Citation Analysis）三部分。Citation Rankings 可提供 22 个专业领域的科学家、机构、国家、期刊、论文的统计分析及排序；Most Cited Papers 包括近 10 年高频被引文献（Highly Cited Papers）、近两年热点文献（Hot Papers）；Citation Analysis 由引文分析基线（Baselines），研究前沿（Research Fronts）两部分构成，其中 Baselines 包括以图表显示的 22 个学科 1994—2004 年间文献平均被引率（Average Citation Rates），以图表显示的 1994—2004 年间某学科分别在 0.01%，0.1%，1%和 10%水平上被引期刊的数量（the Percentiles Table）。Research Fronts 是通过聚类分析定义的特定领域内高频被引的论文集合。为了提高 ESI 中图、表及数据的可读性，Analytical Tools 设立了评述栏目（Commentary），包括 In-cites、Science Watch，主要介绍经典引文注释以及被收录的科学家、研究机构、国家、期刊以及论文的背景，Special Topic 主要为 ESI 筛选出的最新研究领域提供引文分析和专家注释。因此，ESI 为科学研究者提供了一种动态的、综合的、基于 Web 的研究分析环境。

2004 年 ISI 推出的新版 Web of Knowledge 在分析与检索功能以及个性化服务方面皆有所发展。依托 SCI 独特的引文机制、凭借 Web 链接特性，Web of Knowledge 将不同形式、不同来源的信息联系在一起，形成一个知识整合平台，为检索、分析、组织学术信息提供了一种资源。

1.2 NEC 的 Citeseer

Citeseer 于 1998 年由美国普林斯顿大学 NEC 研究院开发，其创建者是 S. Lawrence 等，又名 ResearchIndex。Citeseer 是在自动引文索引（Autonomous Citation Indexing，ACI）机制基础上建设的一个学术论文数字图书馆[5]，它提供了一种通过引文链接检索文献的方式。ACI 能够自动从网上电子形式的文献中建立引文索引，并对文章进行剖析，抽出引文，从而鉴别出引文的上下文，其中包含全文及引文索引，并支持关键词与引文链接检索。目前在其数据库中可检索到超过 50 万篇论文。主要涉及计算机科学领域[6]，概括起来，Citeseer 具有下述功能：

1）支持布尔检索，能够检索出相关文献，浏览并下载网上 Postscript 和 PDF 文件格式的学术论文。检索中可限定检索词的位置，如：要求检索词在标题中（Header）、题目中（Title）等，检索结果可按照年引文量（Citations by Year）、引文量（Eexpected Citations）、标引日期（Date）等排序。

2）浏览并下载该文的原文，同时给出该文的文摘。

3）查看某一具体文献的“引用”与“被引”情况，列出该文主要参考文献、每条参考文献被引频次、参考文献在来源文献中的上下文，绝大部分可得到全文。同时通过引文链接可以获得该文后继文献引用的信息，并可直接看到引文的上下文，绝大部分可得到全文。进而所有引文都可以继续查看其引用与被引情况，因而具有延伸与扩展能力。

4）图表显示文献被引量的时间分布。可依此推测学科热点和发展趋势。

5）查看某一文献的相关文献（Related Records）。Citeseer 采用三种算法计算文献相关度，即基于向量空间的 TFIDF（词在文章中的词频与该词在文集中篇幅之比）算法；通过文章标题矢量距离比较发现相似的标题以及 CCIDF（Common Citation Inverse Document Frequency）算法，用于发现具有相同引文的文章。同时，该系统也提供同被引检索相关文献。

1.3 南开普顿大学的 Citebase

Citebase 是始于 1999 年为期 3 年的开放引文计划（Open Citation Project）的产物，是英国南开普顿大学引文分析专家以及美国康奈尔大学数字图书馆数据管理专家合作之结晶。作为以网络引文分析与引文检索为目的的服务工具，依据文献的影响排列检索结果，因而被誉为学术论文之“Google”[7]。其主旨面向网上公开获取的论文，提供电子印本库的参考与链接服务，与此同时提供引文与影响分析。

当一篇新文章以电子印本的形式存储或发表，Citebase 将自动检索并在搜索引擎中标引，并链接其参考文献，继而在 Citebase 中生成一个该文的 OAI（Open Archives Initiative Protocol for Metadata Harvesting）标识符。系统支持依据引用文献的作者、提名、文摘关键词、出版物名称（Publication Title）、创建日期（Creation Data）以及 OAI 识别号检索，并可按照创建日期、最新更新日期、论文被引量、作者被引量、作者点击率，文章点击率等多种准则排列检索结果。其中，点击率涉及 1999 年 8 月至今的数据，并仅限于英国 Arxiv 镜像站的资料。并采用图示显示文献被引率、点击率随时间的发展情况，因而可直观判断文献的影响及研究热点。应该说明的是，目前 Citebase 采用英国 UTF-8 编码框架，结果为精确匹配。

对于每一篇文献，Citebase 提供了：①文章引用或点击历史；②该文参考文献列表；③引用该文的前 5 篇文章（依文献被引量排列，亦可选择显示引用该文的所有文章）；④与该文同被引的前 5 篇文章（亦可选择与该文同被引的所有文章），并且通过链接，由一篇文献可无限延伸下去，充分体现了引文索引滚雪球式的扩展功能。

目前，该系统主要面向物理学家提供服务，主要资料

来源于康奈尔大学的 Arxiv 电子印本服务系统，其中包括1991年以来的物理、数学、计算机科学的 20 万篇文献；认知科学电子档案（Cogprints），其中包括心理学、神经科学、语言学和计算机科学的 1 400 篇文献；生物医学中心出版社（BioMed Central）提供的生物医学研究论文 900 篇。Citebase 现有记录 23 万篇，标引参考文献 560 万篇。

该系统目前处于试验阶段，其主页郑重声明：本系统正处于试验阶段，使用者请小心从事，目前尚不能用于学术评价。为了对其可行性（Usable），有益性（Useful）做出评定，2002 年 6—10 月南开普顿大学对来自不同背景的 200 个用户进行了调查。结果表明：用户认为该系统使用简单，设计合理，尤其采用的同被引的设计思想在引文索引史上是一个创举，但与此同时，用户也提出了应该扩大收录范围、提高检索查准率的建议。

2 以 Web 为基础的引文索引体系比较

2.1 设计思想同出一辙

1955 年，E. Garfield 受到 1870 年出版的法律工具书——《谢拨德引文》的启发，开创了科学引文索引。其初衷是从引文的视角标引和组织科学文献，以利于文献的检索与利用，现今已经发展成为学术界公认的科学技术文献检索的工具。事实上，网络环境下出现的 Citeseer、Citebase 均是建立在引文索引基本原则的基础上，正如 Citebase 评估报告中所陈述的：Citebase 按照引文影响排列并提供检索服务。Citeseer 在其主页郑重声明：Citeseer 利用 ACI 自动产生引文索引以利于文献检索和评估。所以，就设计思想而言，二者完全拷贝了 SCI 的思想精髓[8]。因而也都具备了超时空、跨学科组织文献的特点。

2.2 检索功能不断完善

作为检索工具，三种引文索引借助于引文编织科学论文网络，并不断进行自我完善。新版的 Web of Knowledge 主页更为直观，布局设计简单，提供了更多的检索选项，突出个性化服务特点，检索功能一目了然。检索结果上限由最初的 500 个增至 10 万个，增加了 DOI 标识，因而在原有 800 万篇基础上增加了 1 000 万篇全文链接，并与 250 家出版社建立了协作关系，其中包括美国化学协会的各种期刊，因而与出版商的协作增至 300 家。此外，实现了 Openurl 链接，增加了团体作者检索。并预计于 2005 年推出新的产品——the Century of Science，这一计划将使 Web of Science 的功能回溯到 1900 年，届时被引文献检索与导航年限为 1900 年到 2004 年。Citebase 也在积极进行性能改进并扩大影响，全文检索从 3 秒下降到 1.5 秒，Web 使用统计表明，评估中日点击率从最初的 25～45 次上升到 660 次，尤其在物理学家中扩大了影响。Citeseer 除了提供多途径检索之外，重点突出了来源文献的上下文。不仅如此，基于 Web 的非线性结构将引文索引关联到整个文献检索系统中而建立的引文索引，充分发挥了链接的优势，强化了其信息提供的功能。Web of Knowledge 跨数据库检索整合不同类型资源，与不同出版社网络版全文链接，提供从网上直接获取全文；而 Citeseer 和 Citebase 直接从网上获取公开发表的学术论文。

2.3 分析功能不断强化

作为 Garfield 设立引文索引的副产品，引文分析的功能引起了格外的关注。通过引文追溯文献之间的内在联系，就可以找到一系列内容相关的文献以及某一研究领域、某一学术观点的发展脉络，从而可以看出某一学科或领域的研究动态和发展趋势，并根据某一学术概念、某一方法、某一理论的出现时间、出现频次、衰减情况等，分析出学科或领域研究的走向和规律。同时还逐渐发展成为评价一个国家、一个地区、某个单位以至个人科研成果及其学术影响的极为重要的工具之一，甚至作为引文分析的重要方法之一——同被引已经远远超出了情报检索的原则，引文索引也远远超出了其传统内涵。这一点就连 Garfield 本人也是始料未及的。Web 环境下的引文分析资源更为丰富，分析功能得到了大力的发展。目前，三个系统均采用图表形式显示影响与趋势。新版的 Web of Knowledge 融合了诸多分析功能，用户可以迅速分析和组织检索结果，从时间、机构、学科、作者等选项更深入了解检索结果；Citeseer 显示某一主题文献（或某一作者、机构所发表文献）的时间分布；Citabase 则是从揭示文献作者的影响为起端。从目前来看，Web of Knowledge 揭示的内容宏观与微观相济，后者则相对微观。

2.4 分析方法各有千秋

相关文献检索是研究人员了解其研究主题发展脉络、最新动向不可或缺的重要渠道。目前三个系统均具备此项功能。只不过 Web of Knowledge 采用的是引文耦合的方法，而 Citabase 采用的是同被引的思想，Citeseer 则采用计算机的算法，并提供同被引检索相关文献。从引文分析的角度而言，1973 年 Small 开发的同被引技术是 1963 年 Kessler 提出的文献耦合概念的创新和逆向思维的发展，并且近年来信息可视化及网络寻址定标技术的发展使同被引技术焕发了新的活力，美国费城德瑞克赛大学基于同被引新近开发的 Authorlink 系统被誉为知识信息提供服务的一次革命[9]。因而就此项功能而言，Citebase、Citeseer 要比 Web of Knowledge 更为科学，并且更具发展潜力。

2.5 收录范围互为补充

科学引文索引囊括了自然科学、社会科学、艺术与人文科学诸领域的最权威期刊。仅 SCI 即收录了自然科学、

工程技术、生物医学等150多个学科领域内的核心期刊5 800多种，而目前上述两个系统尚处于试验阶段，Citeseer仅包括计算机科学的文献，而Citebase涉及物理学、数学、计算机科学、生物医学、心理学等，目前主要的服务对象是物理学家。就数量和学科范围而言，无法与Web of Knowledge相提并论。扩大收录范围是包括物理学家在内的被调查者对Citebase提出的主要意见。因此，Citebase近期计划收录包含近30个国家100多种经济类期刊、研究报告、软件以及信息量超过207 000条的公共数据库RePEc（Research Papers in Economics）和电子印本数据库（eprints. org Repositories）。但是，客观而言，在收录的情报源类型方面，开放环境下的Citeseer以及Opcit的产物Citebase一定程度上是美国科学情报研究所编辑出版的SCI的补充。由于SCI受到手工劳动的限制，并以期刊文献为主，且一直将焦点聚集在高质量的期刊，虽增设了会议录，但仍有专著、技术报告、预印本等科学交流的重要情报源被排除在外，这种选择性标引的局限性是不言而喻的，因而基于SCI数据所做出的评价也就难免有失偏颇，这也是对SCI批评由来已久的主要原因。而Citeseer和Citebase充分发挥了Web的优势，将收录范围拓展到开放环境。Citebase贯彻了2001年12月1—2日在布达佩斯召开的Open Access Initiative会议精神：开放存取，帮助学者发现与其研究相关的文献。而Citeseer的宗旨就在于有效地组织网上文献，多角度促进学术文献的传播与反馈，因此开放环境下的引文分析更为准确、快捷。

2.6 服务性质相差悬殊

首先，具有50年历史的SCI及新近改版的Web of Knowledge服务体系完善，设计思路清晰，结构合理。相对而言，Citeseer则显得有些凌乱，使人费解。Citebase设计简单，可信性强。究其原因，SCI出自引文分析学家之手，Citeseer出自计算机学家之手，Citebase则是综合了计算机学家与引文分析学家的优势。所以，在设计思路方面有所差异也就在所难免。其次，就标引深度而言，Web of Knowledge标引核心期刊的题名，而Citeseer和Citebase索引网上自由获取的文献的全文。第三，服务性质方面，Web of Knowledge价格昂贵，而目前Citeseer和Citebase是免费的。免费服务的实施将促进引文分析思想的普及，其所设立的引文索引也将为更多的人所采用，从而有利于学术论文的传播。

3 结论

从上述三个系统功能及其比较可以看出：

1）三个系统相互借鉴而发展[10]，诚然，与具有50年历史的SCI相比，Citeseer和Citebase尚处于初生期，但就Web of Knowledge 2004年的大手笔及其2005年的新计划来看，后者的出现对其有所触动。可以这样认为：新型索引借鉴了SCI的思想，反过来，新索引的出现促使Web of Knowledge的改进。

2）以Web为基础具有引文链接机制的新型引文索引的发展将有效地促进科学发展。作为学术信息资源发现的工具，已经不是传统意义上的引文索引，而是一个通往许多其他信息数据库，包含丰富信息资料的强大浏览工具。引文索引及其引文链接为研究人员提供了发现、检索、浏览、借鉴前人工作成果的空间。

3）新一代的引文系统更加突出引文分析功能。具有强大引文分析功能的新型引文索引的发展将有效地促进科学交流。Crossref等参考文献链接系统证明了Web环境下出版商前所未有的联合，但是作者们并未对开放存取赋以极大的热情，问题的关键就在于不具备完善的评价作者影响（Impact）及明显度（Visibility）机制。因此，从长远来讲，以Web为基础具有强大引文分析功能的引文索引系统的发展将更加有效地促进科学交流。

4）Open Journal引发了Web of Science的出现，Open Citation Project促进了Citebase的出现，所有的一切都是在开放存取、促进学术信息资源的传播的机制下发生的。因此，开放存取、参考文献链接、引文分析是相互关联的。Web of Knowledge走向开放环境将是大势所趋。□

参考文献

1 王崇德. 文献计量学教程. 天津：南开大学出版社，1996

2 Cronin B. Bibliometrics and Beyond: Some Thoughts on Web-based Citation Analysis. Journal of Information Science, 2001 (1)

3 莫梅琦，张嵬. ISI Web of Knowledge体系检索特色与应用评析. 现代图书情报技术，2003 (1)

4 http: //www. isiwebofknowledge. com/whatsnew/

5 Lawrence S, et al. Digital Libraries and Autonomous Citation Indexing. IEEE Computer, 1999 (6)

6 http: //citeseer. ist. psu. edu

7 http: //citebase. eprints. org/cgi-bin/search

8 Hitchcock S. Open Citation Linking. D-lib Magazine, 2002 (10)

9 White H D. Pathfinder Networks and Author Cocitaiton Analysis: A Remapping of Paradigmatic Information Scientists. Journal of the American Society for Information Science and Technology, 2003, 54 (5)

10 张晓林. 开放环境下的参考文献链接. 现代图书情报技术，2002 (1)

作者简介：宋丽萍，讲师，博士生。

收稿日期：2004－12－09