

●冯 群, 朱 静, 黄如花 (武汉大学 信息管理学院, 武汉 430072)

开放存取资源搜索引擎的比较与分析

【关键词】开放存取; 搜索引擎; Google Scholar; Scirus; OAIster

【摘 要】对 Google Scholar、Scirus 和 OAIster 这 3 个综合性的开放存取资源搜索引擎的资源来源、检索功能、全文免费获取情况进行详细比较, 分析其不足, 提出了改进的建议。

【中图分类号】G250.72

【文献标志码】B

【文章编号】1005-8214(2009)08-0018-03

开放存取 (Open Access, OA) 资源的数量不断增加、质量不断提高, 对图书馆扩大馆藏资源非常重要, 逐渐引起了图书馆界的重视。但是, OA 资源分散多处, 其登记也缺乏一个完整的机制, 因而, 目前对其获取主要借助于 OA 资源的搜索引擎。^[1] 本文将对 Google Scholar、Scirus、OAIster 这 3 种综合性学术搜索引擎在 OA 资源存取方面进行比较, 以期分析其各自的优势与不足, 提出优化建议。

Google Scholar 是 Google 的众多产品之一, 作为综合性的学术搜索引擎, 免费为大众提供学术文献搜索服务。^[2] Scirus 是一款由 Elsevier 科学出版集团推出的免费的学术搜索引擎, 是最全面的网络科技专用搜索引擎, 为用户提供期刊检索服务, 同时免费提供网络信息资源的检索服务。^[3] OAIster 是密西根大学图书馆数字图书馆建设服务部的一个项目, 作为一个数字资源联合目录, 通过利用开放档案创始计划元数据采集协议 (Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH) 采集数字资源的元数据, 提供对来自 1005 个资源提供者的 1600 余万条记录的检索, 能够检索到包括文本、图像、音频、视频和数据集等多种类型的资源。^[4]

1 三种搜索引擎收录范围的比较

收录信息资源的范围对于搜索引擎的检索效果有着重要的影响。下面将从收录的数量、学科范围、资

源格式或类型、资料来源、资源的语种 5 个方面对这 3 种搜索引擎进行比较。^[5]

1.1 资源的数量

3 个搜索引擎资源数量分别是 Google Scholar 大于 15.5 亿网页, Scirus 有 4.5 亿网页, OAIster 有 16973216 条记录, Google Scholar 显然占了很大的优势。

1.2 资源的学科范围

Google Scholar 和 Scirus 覆盖的学科范围都比较广, 二者都提供了根据学科范围来限定检索范围的检索功能。在能够检索的范围之内, 以自然科学领域为主, 人文艺术、社会科学领域比例较少, Google Scholar 为 28.6%, Scirus 也只有 31.8%, 并且领域分类也比较含糊。而 OAIster 没有统一的学科分类, 各个资源提供者各自分类, 相对较混乱。

三者都不提供按学科来浏览资源的功能, 不能满足用户从学科范围获取相关的 OA 资源的需求。

1.3 资源格式或类型

Google Scholar 以论文检索为主, 与专门检索图书的 Google book 相互补充。Scirus 则可以根据检索到的资源格式进行再次筛选检索结果, 但是其结果显示首先是 HTML 网页, 不符合大多数用户以 PDF 格式为首选的要求。

1.4 资源来源

Google Scholar 利用信息采集器 (crawlers) 采集信息, 经过“学术性”要求的筛选, 将诸如商业杂志等不符合要求的排除在外, 其他信息均能够被 Google Scholar 搜索到。这些信息的格式可以是 PDF 等各种格式, 且鼓励提供全文, 不收录按钮或图片。如果图书馆想让其资源被 Google Scholar 搜索到, 就必须是 OCLC Open Worldcat program 的成员。

Scirus 主要收录期刊论文和 Web 资源, 前者来自与之合作的 23 家出版商, 后者包括科学家个人主页、新闻网页、政府网站、社区主页、公司网站、会议信息等。

OAIster 的资源主要是来自与其合作的 1005 个资

【基金项目】本文系教育部“新世纪优秀人才支持计划 (NCET) 资助” (教技司 [2007] 209 号) 成果之一

源贡献者,不从其他学术网站或者与学术相关的网站上采集资源。

可见, Google Scholar 的资源来源最为广泛, Scirus 次之, OAIster 最少。除了 Google Scholar 外, 后两者都向公众公布了他们的具体的资料来源。另外, 由于 Google Scholar 只面向文本, 后两者的资源格式还包含了图片或者音频视频等, 其响应速度明显慢于 Google Scholar。

1.5 资源的语种

3 种搜索引擎的资源都以英文为主, 其中当属 Google Scholar 收录的资源语种最为广泛, 包括德文、日文、法文、英文、葡萄牙文、西班牙文、韩文、简体中文、繁体中文, 等等。用户还可以在“使用偏好”中设置自己要检索的资源语种, 中文版还可以在基本检索时就进行简体中文、繁体中文或者是全部网页的选择。

三者都收录有中文(简体中文和繁体中文)资源, 并且支持简繁体中文的自动转换。以 Google Scholar 能够检索到的中文资源数量最大, 但是总体看来三者的中文资源相对于各自的整体资源来说所占比例极低。

2 三种搜索引擎的检索功能比较

检索功能对于一个搜索引擎来说至关重要, 能否更加简洁方便地检索到更多的、更相关的开放存取资源是搜索引擎的重要的评价指标。

2.1 基本检索

3 个搜索引擎都支持布尔逻辑检索、短语检索、截词检索, 都不支持邻近检索, 不区分大小写。略有不同的是, Scirus 除了支持词尾的截词检索(*)外, 还支持中间截词(?)。

2.2 高级检索

Scirus 的检索功能较为强大, 字段限制、时间限制、资源类型限制、文件格式限定以及资源来源限定都十分详细具体, 可选项较多, 学科限定包括 22 个学科。^[6]

Google Scholar 的高级检索有字段限制标题和整篇文章、作者检索、出版物限定(用户自己输入)、时间限定、学科限定 7 大类 21 个学科。^[7]

OAIster 的检索功能最弱, 高级检索和基本检索同属一个页面。仅有 3 个检索词输入框, 一个字段限制, 包括标题、作者或者责任者、学科、语言、整个记录, 一个结果排序方式选择, 包括权重词出现频率(默认)、标题、作者、时间、检索词出现频率等等。^[8]

2.3 使用偏好

使用偏好设置能够在一定程度上限制检索范围。Google Scholar 的“使用偏好设置”包括接口语言和检索语言设置、图书馆链接设置、每页显示结果数量、是否在新的窗口打开检索结果、是否显示引用链接。Scirus 的“使用偏好设置”包括每页显示的数量、是否在新的窗口打开检索结果、是否按领域聚类检索结果、通过选择资源来源机构限制检索范围。

2.4 自动纠错功能

Google Scholar 和 Scirus 都有自动纠错功能, 当笔者输入 anniline 时, 在检索结果页面后会有“Did you mean: aniline”的提示, 但是 OAIster 没有自动纠错功能。

3 三种搜索引擎的检索结果可开放存取情况的比较

虽然这 3 个搜索引擎提供的检索服务都是免费的, 都能够检索到大量的资源, 但这些资源能否免费获得, 免费获得的方式是否复杂等关于开放存取的问题对用户的影响更为重大, 也是评价 OA 学术资源搜索引擎的重要指标。

笔者采取对比实验的方法, 从检索结果数量、免费获取全文的比例、收费获取全文的比例等方面对它们的检索效果进行了简单的测评。即选取含义明确的词汇(search engine)作为检索, 分别在 3 个搜索引擎的基本检索中进行检索, 同时选取前 30 个结果作为测评依据。(限定为 PDF 格式)

表 以“search engine”为关键词进行检索的检索结果开放存取情况比较

	Google Scholar		Scirus		OAIster		
检索结果数量(条)	1,970,000		5,211,560/16,933		3293		
免费获取全文数量及比例	22	3(13.6%)	24	13(54.2%)	23	0	1步
	73.3%	19(86.4%)	80%	11(45.8%)	76.7%	22(95.7%)	2步
		0		0		1(4.3%)	3步
收费获取全文数量及比例	3	10%	0	0%	0	0%	
无全文获取途径数量及比例	1	3%	4	13.3%	7	23.3%	
无链接或死链接数量及比例	4	13.3%	0	0%	0	0%	
不相关的数量及比例	0	0%	2	6.7%	0	0%	

(注: 表中 1、2、3 步分别指需要经过 1、2、3 个步骤能免费获取全文)

总体上说, 检索并免费获得 OA 资源效果较为令人满意, 都达到了 70% 以上, Scirus 更是达到了 80%。Google Scholar 和 OAIster 检索到的 OA 资源绝大多数要

经过两个步骤才能够获取全文, Google Scholar 经过两步获取全文占能够免费获取全文的 86.4%, 而 OAIster 则高达 95.7%, 这一数据正好和 3 个搜索引擎资源来源相吻合。OAIster 的资源主要来自 1005 个资源贡献者, 要获得 OA 资源的全文, 需通过检索结果页面上的相关链接进入资源提供者的页面, 然后才能够通过下载或其他方式获取全文。Scirus 除了有固定的资源贡献者外, 还广泛收集网上的学术信息, 能够一站获取的资源比例相对较大, 但 Scirus 的检索结果中并没有完整的摘要介绍, 因而很可能在 1 步点击下载后得到一些不相关的文献或者只是摘要而非全文。若能把 OAIster 结果显示页面的“主题”融入进来将会更有利于用户方便快捷地获取相关的 OA 学术资源。

Google Scholar 收费资源占到了 10%, 对于中文文献来说比例更大, 对用户来说是极不方便的, 而收费获取全文在另外两个搜索引擎中并未出现。要指出的是, 并不是搜索引擎要收费, 而是资源来源机构要收费。如果能够在结果显示页面对收费与否的记录加以区分, 用户的使用将会更加方便。

有些资源贡献者只提供摘要信息并不提供全文, 使得 OAIster 的无全文获取途径比例最高。Google Scholar 的比例低与之广泛的资源来源分不开, 无链接或死链接比例较大也跟其资料来源太广密切相关。

4 三个开放存取资源搜索引擎的特色与不足

4.1 Google Scholar

Google Scholar 的响应速度非常快; 资源非常广泛, 无论是在资源的语种还是资源的学科范围上都有很大的优势; 检索界面和检索结果页面摆脱了单调的页面颜色, 视觉效果很好; “相关文章”和“被引次数”为用户全面了解某一学科主题或其最新进展提供了便利; 自动推荐作者的功能为对某一领域不甚熟悉的用户提供了很大的帮助。另外, 针对中文用户还推出了 Google Scholar 中文版。

当然, 其高级检索功能还有待提高, 不能根据资源的类型或者格式来缩小检索范围, 使检索结果的排序方式仅有一种默认的排序方式, 给用户的信息筛选造成了不便。Google Scholar 在 OA 资源全文的免费获取上存在较大问题, 一旦信息采集器被允许进入, 大量资源就会被 Google Scholar 检索到并提供给用户, 但又无法保证资源提供的有效性, 用户检索到资源却无法免费获取利用, 这一点在中文文献上尤为突出。^[9]

4.2 Scirus

Scirus 突出的特色是检索功能强大, 不仅具有齐

全的高级检索功能, 并对检索结果进行各种分类来缩小检索范围提高检准率。虽然其资源数量不及 Google Scholar, 但资源格式丰富, 如果用户需要获取书籍、图片及相关网页等多种资源, Scirus 是个不错的选择。它还能够为检索到的结果提供保存、E-mail, 导出为文本文件的处理。

一般来说图片等格式会影响检索速度, 因而 Scirus 的响应速度较慢。在排序上有选择但过于简单, 只有按相关度和时间排序。Scirus 的检准率在 3 个搜索引擎中相对较弱, 尤其是在中文资源的检索上, 很多不相关的信息都会收录。因此, 用户想要通过 Scirus 准确地查找学术信息一定要先熟悉它的各种检索方法才能缩小检索范围。^[10]

4.3 OAIster

OAIster 的检索结果排序方式多样, 可以满足不同用户的不同需求, 但它需要用户在检索时就进行选择, 一旦确定就不能在检索结果页面重新选择。对检索结果的整合比较少, 无确定的主题分类规则, 给用户带来了不便。OAIster 的 OA 资源在三者中是最少的, 通过它只能检索到与它合作的资源贡献者的资源, 限制较大, 也正因为如此, 检索结果的相关度高成为了 OAIster 的一大优势。

5 优化三大开放存取资源搜索引擎的建议

5.1 扩大资源收录范围

(1) 学术性。这是学术性搜索引擎区别于其他搜索引擎的关键之处。开放存取运动的目的是为了促进学术信息交流, 开放存取资源搜索引擎理应加强学术信息的收录。

(2) 广泛性。表现在学科范围、资源格式与类型以及语种三个方面。学科范围上, 尤其要加强人文、艺术、社科类资源的收集; 资源的格式和类型上, 以保证搜索引擎的检索速度为前提, 丰富格式类型; 语种上, 尤其要加强中文免费学术资源的搜集。此外, 搜索引擎应加强和各类学术机构的合作, 例如大学、科研机构 and 图书馆等; 同时要加强对网上免费的学术资源的搜索。

(3) 开放性。这里的开放性主要是指搜索引擎搜集到的信息是否能够免费获取, 即便不能, 应在检索结果中增加一定的标志, 从而避免用户经过多个步骤后才发现资源不能免费获取。

5.2 增强检索功能

(1) 使检索界面更加人性化。关于如何使用该搜索引擎以及帮助的标志应该放在搜索(下转第 24 页)

态,它在不断向着有序化程度更高的水平和层级演化,目前维基百科的注意力已经从关注增长转移到关注质量问题上来。

[参考文献]

- [1] Pierpaolo Dondio, Stephen Barrett. Computational Trust in Web Content Quality: A Comparative Evaluation on the Wikipedia Project [EB/OL]. [2008-06-06]. [http://www.informatica.si/PDF/31-2/02_Dondio Computational % 20Trust % 20in % 20Web % 20content % 20quality...pdf](http://www.informatica.si/PDF/31-2/02_Dondio%20Computational%20Trust%20in%20Web%20content%20quality...pdf).
- [2] Jakob Voss. Measuring Wikipedia [EB/OL]. [2008-06-18]. <http://eprints.rclis.org/archive/00003610/>.
- [3] 维基百科可靠吗? [EB/OL]. [2008-05-20]. <http://www.sciencevie.cn/gb/article/20085201622066190.htm>.
- [4] WikipediapageHistory [EB/OL]. [2008-05-20]. <http://en.wikipedia.org/wiki/Http:Pagehistory>.
- [5] BesikStvilia, etc. SmithInformationQualityDiscussions Wikipedia [EB/OL]. [2008-06-20]. <http://mail-er.fsu.edu/~bstvilia/papers/qualWiki.pdf>.
- [6] WikipediaTalkPageGuidelines [EB/OL]. [2008-05-20]. <http://en.wikipedia.org/wiki/Wikipedia:Talkpageguidelines>.
- [7] Wikipedia Featured Articles [EB/OL]. [2008-05-20]. <http://en.wikipedia.org/wiki/Wikipedia:Featuredarticles>.
- [8] Wikipedia Featured Articles Candidates [EB/OL]. [2008-05-20]. <http://en.wikipedia.org/wiki/Wikipedia:Featuredarticlecandidates>.
- [9] Wikipedia Featured Articles Criteria [EB/OL]. [2008-05-20]. <http://en.wikipedia.org/wiki/Wikipedia:Featuredarticlecandidates>.
- [10] Wikipedia Articles Delection [EB/OL]. [2008-05-20]. <http://en.wikipedia.org/wiki/Wikipedia:Artoclesfordelection>.

[作者简介] 王丹丹 (1980—), 女, 河南科技大学讲师, 中国科学院国家科学图书馆博士研究生, 研究方向: 网络化信息服务, 已发表论文 6 篇。

[收稿日期] 2008-11-05 [责任编辑] 陈永平

(上接第 20 页) 引擎首页的醒目位置, 帮助用户快速地掌握该搜索引擎的检索技巧。同时高级检索提供检索字段应该尽可能地满足用户的需求。

(2) 加快检索速度。本文的 3 个搜索引擎中, Google Scholar 的检索速度最快, 其他两个搜索引擎的检索速度还有待提高, 检索速度过慢很可能引起用户的反感而被放弃使用。

(3) 检索结果显示清晰。检索结果应该显示摘要、被引情况、提供被引的链接、所在主题或学科(最好能够链接到相应主题或学科的其他资源)以及是否能够免费获取的标志, 提供多种检索结果排序方式方便用户的查找。

综合性的学术搜索引擎如果能够努力做到以上几点, 将会给用户利用免费学术资源带来极大的便利, 同时促进开放存取运动的发展。

[参考文献]

- [1] 李春旺. 网络环境下学术信息的开放存取 [J]. 中国图书馆学报, 2005 (1): 33-37.
- [2] Google 学术搜索引擎·关于 [2009-03-11]. <http://scholar.google.com/intl/en/scholar/about.html>.

- [3] Scirus 搜索引擎·关于 [2009-03-11]. <http://www.scirus.com/srsapp/aboutus/>.
- [4] Oaister 搜索引擎·关于 [2009-03-11]. <http://www.oaister.org/about.html>.
- [5] 蒋亚琳. 对三种学术搜索引擎的析评 [J]. 情报探索, 2007 (1): 46-48.
- [6] Scirus 搜索引擎·帮助 [2009-03-11]. <http://www.scirus.com/html/help/index.htm>.
- [7] Google 学术搜索引擎·高级检索指南 [2009-03-11]. <http://scholar.google.com/intl/en/scholar/help.html>.
- [8] Oaister 搜索引擎·帮助 [2009-03-11]. <http://www.oaister.org/help.html>.
- [9] 程妮. 科学搜索引擎 Scirus 研究 [J]. 现代图书情报技术, 2005 (3): 45-49, 52.

[作者简介] 冯群, 女, 武汉大学信息管理学院图书馆学系 2006 级本科生; 朱静, 女, 武汉大学信息管理学院图书馆学系 2006 级本科生; 黄如花, 女, 武汉大学信息管理学院教授、管理学博士、博士生导师。

[收稿日期] 2009-03-16 [责任编辑] 陈永平