·用户服务与研究·

增强机构知识库内容发现和利用影响的策略与方法实践*

卢利农 祝忠明 张旺强 刘 巍 姚晓娜
(中国科学院国家科学图书馆兰州分馆 甘肃兰州 730000)

摘 要:文章以中国科学院机构知识库 CAS OpenIR 为例,采用基于学术搜索引擎 Google Scholar 优化的策略和方法,如针对 Google Scholar 收录原则、Google Scholar 元数据体系、sitemaps、Robots 协议等策略和方法进行分析和实践,通过提升机构知识库在 Google Scholar 中的收录比率,进而增强机构知识库中内容被发现引用的机率,以扩大 IR 利用影响力。

关键词:机构知识库 谷歌学术搜索 学术搜索引擎优化 中科院机构知识库
中图分类号:G252 G255.76 **文献标识码**:A **文章编号**:1003-6938(2012)05-0085-05

Strategies and Methods to Improve IR Discovery and Influence

Abstract Taking CAS OpenIR institutional repository of the Chinese Academy of Sciences as an example, the authors tries to improve, mainly based on the ASEO strategy and method, the IR's index ratio in Google Scholar and further strengthen the influence of the quotes and use of the content of the institutional repository. This paper mainly analyzed Google Scholar Principles of Collecting, Google Scholar Metadata system, Sitemaps, Robots etc. The ASEO research and practice process also play a positive reference role for other digital library system.

Key words IR; Google Scholar; ASEO; CAS OpenIR

1 引言

近年来机构知识库(Institutional Repository,IR)快速稳步增长,已覆盖了大部分知名高校和科研机构。目前在开放获取机构资源库 OpenDOAR 中注册登记的 IR 已有 2163 家[1],除此以外还有相当一部分数量的 IR 未在 OpenDOAR 中注册。IR 做为支持开放获取的一种重要形式,支持机构实施数字知识资产的长期保存和管理,提高机构及科研人员智力成果的发现几率、传播范围和影响,是吸引机构及科研人员重视和参与 IR 建设的重要因素。相关的研究也表明,支持开放获取的论文其引用影响可获得 25%~250%的提升[2]。而 Arlitsch 等人[3]的调查结果显示,当前 IR 内容被 Google Scholar 收录的比率总体上维持在 10%~30%的水平,甚至有 0%的 IR(见图 1)。也就是说,大部分 IR 的内容没有得到充分的发现和利用,仍然局限在小范围内进行交流传播。

Google Scholar 作为一项针对学者和科研人员的免费学术文献搜索服务,现在已成为学者、研究人员和学生查找专业文献资料的首选工具[4]。其搜索的范围涵盖了几乎所有知识领域的高质量学术研究资料,包括论文、专业书籍以及技术报告等。Google Scholar 不但可以过滤普通网络搜索引擎中对学术人士无用的大量信息,通过与众多学术文献出版商的合作,还加入了许多普通搜索引擎无法搜索到的内容。目前,科研用户通过网络来获取资源,第一选择就是通过 Google 等搜索引擎进行大范围搜索,其次考虑利用专业的学术数据库,最后才会去翻阅学术期刊。这种检索顺序已经形成了一种社会习惯。

因此,如何解决 IR 被搜索引擎 Google Scholar 收录,提升 IR 中学术文章被 Google Scholar 收录的比率,已成为增强 IR 内容可发现性和可见性的关键。本文以中国科学院研究所 IR 平台 CAS OpenIR[5]为例,采用学术搜索引擎优化(Academic Search Engine Optimization,ASEO)的策略和方法,通过提升 IR 在 Google Scholar 中的索引比率,进而增强 IR 中内容被发现引用和利用影响力。

* 本文系中科院知识创新工程重要方向项目"研究所机构知识库建设"和中科院西部之光联合学者项目"机构知识库的语义增强方法与技术研究"研究成果之一。
收稿日期:2012-08-07 责任编辑:魏志鹏

Repository Name	Repository Software	Repository URL	Repository items	Items in Google Scholar	Indexing Ratio
Boston College-eScholarship@BC	DigiTool	dcollections.bc.edu	1,635	1	0%
UW - ResearchWorks Archive	DSpace	digital.lib.washington.edu/dspace	11,285	304	3%
Univ of Rochester Research	IR+	urresearch.rochester.edu	16,184	983	6%
CaltechAuthors	Eprints	authors.library.caltech.edu	22,000	2,290	10%
D-Scholarship@Pitt	Eprints	d-scholarship.pitt.edu	5,888	686	12%
Columbia Univ-Academic Commons	Digital Commons	academiccommons.columbia.edu	4,631	586	13%
IU Scholarworks	DSpace	scholarworks.iu.edu/dspace	7,782	1,030	13%
Texas A&M Repository	DSpace	repository.tamu.edu	46,324	7,250	16%
UW Madison-Minds@UW	DSpace	minds.wisconsin.edu	15,078	2,520	17%
eCommons@Cornell	DSpace	ecommons.library.cornell.edu	18,544	3,410	18%
Harvard Univ - DASH	DSpace	dash.harvard.edu	6,193	1,710	28%
Univ of Oregon-Scholars Bank	DSpace	scholarsbank.uoregon.edu/xmlui	9,740	2,840	29%

图 1 IR 被 Google Scholar 收录情况调查表[3]

2 ASEO 策略和目的

ASEO 建立在传统的 SEO[6]基础之上，是从普通的 SEO 发展而来。由于学术搜索引擎 Google Scholar 与普通搜索引擎有着明确的定位区别，因此 ASEO 与SEO 有着明显的不同之处。

SEO 指通过采用易于搜索引擎索引的合理技术手段和策略，使网站各项要素适合搜索引擎的检索原则，从而更容易被搜索引擎收录和优先排序。SEO 基于网页(Web Page)，收录过程较灵活和容易。IR 属于学术产出的数据库平台，有着自身的元数据元素集，其中的学术文章属于“Academic Invisible Web”[7]，不能被 Google Scholar 直接访问和索引。因此，在被学术搜索引擎 Google Scholar 收录前，需要对 IR 进行 ASEO 改造，使其符合 Google Scholar 索引标准，易于被 Google Scholar 收录爬取。即：

(1) 使 IR 可以被搜索引擎 Google Scholar 更好地收录和更新(包括 IR 的元数据和全文)；

(2) 使搜索引擎在规则允许的范围内进行索引，明确 IR 的哪些页面可以被索引收录，哪些页面不能被索引收录；

(3) 在用户使用 Google Scholar 搜索时，可以排名靠前的呈现 IR 中的相关条目，起到推介 IR 的作用；

(4) 将 IR 中开放权限的全文纳入 Google Scholar 的全文检索中，增加 IR 中论文的可见性，提高论文的被引用率。

3 Google Scholar 收录原则和排名算法

Google Scholar 针对学术性数据库内容的收录和索引，有明确的收录原则[8]，如：① 被收录文章需要有唯一的 URL；②匿名用户可免费地通过原文 URL 进入阅读被收录文章；③数据库服务的 Robots.txt 协议正确配置，明确允许及禁止 Googlebot 爬取的路径及内容范围；④数据记录的 Meta 标签符合 Google Scholar Meta 规则，并且必须包含 DC.title，DC.creator，DCTERMS.issued 三项描述元数据；⑤记录除了题录文摘信息外，被收录记录必须要有全文；⑥全文格式为 PDF 格式。

Google Scholar 检索排名继承了普通 Google 检索中应用的 PageRank 算法[9]，即主要看某项学术内容、页面被引用的情况，同时还将文章全文、作者和出版物等因素纳入算法，从而保证检索结果的高相关性，提高查准率。学术论文被引述的频度越多，一般判断这篇论文的权威性就越高，它的 PageRank 值就越高。

4 面向 IR 的 ASEO 策略与方法实现

根据学术搜索引擎 Google Scholar 收录、排名的要约特点，本文中笔者将选取 ASEO 中的关键环节，就设计思路和实现的过程做一分析说明。

4.1 搜索引擎注册

在传统 SEO 过程中，网站管理员不用太担心网站的收录情况，在网站运行一定时间后搜索引擎的机器人会自动通过已被索引的外部链接发现该网站。而学术搜索引擎 ASEO 过程中，往往需要通过管理员在 Google Scholar 中对相关的服务进行注册，来通知机器人将其纳入爬取对象。有鉴于此，在研究所 IR 部署完成后：

(1)要求或者帮助研究所尽快在 Google Scholar 中完成其 IR 的注册和发布。在 Google Scholar 注册 IR 过程中，除了声明 Google Scholar 要求的收录原则外，还需要声明 IR 所用软件、论文数量、语种、访问地址。

(2) 由于 Google Scholar 的 PageRank 算法对网络分类目录尤为重视，如果网站被 ODP(http://www.dmoz.org)、Yahoo! Directory(http://dir.yahoo.com)等网络分类目录收录，则可大幅提升其 PR 值。因此，积极帮助研究所 IR 在重要网络分类目录中进行注册。

(3)随着 OpenROAR(http://www.opendoar.org)、ROAR(http://roar.eprints.org) 等开放知识库注册登记服务在知识库服务领域日益产生重要影响和 Google Scholar 等搜索引擎的合作，我们也应积极引导和帮助研究所 IR 在这些专门性目录服务中进行注册，以加强和提升 IR 被搜索引擎发现和索引的几率。

4.2 建立适合 Google Scholar 发现和索引的描述元标签体系

Meta(网页描述元标签)为 Google Scholar 检索结果的

输出格式提供了基于 DC 元数据标准的标题、作者、出版物名、出版年/期、摘要等内容描述信息。当用户通过 Google Scholar 进行检索时,Google Scholar 自动辨识学术文章的格式与内容,取得描述信息,并针对论文指示的信息建立自动的引用分析。因此,描述元标签及其描述信息十分重要。

要保证 IR 所有内容为 Google Scholar 成功索引,就必须为 IR 所有的记录提供带有 Meta 描述元标签的页面。为此,在 CAS OpenIR 中设计 Meta 标签组为自动生成,不同论文记录页面中的 Meta 值自动从记录对应的内部元数据字段中读取。由于 IR 中条目元数据字段为内部元数据存储字段,并不能直接用于 Meta 标签,因此需要在使用前建立 CAS OpenIR 元数据字段与 Meta 之间的映射关系(见图 2)[10]。

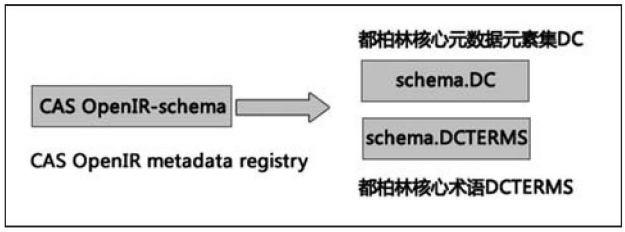

图 2 CAS OpenIR 元数据字段与 Meta 对应关系图

根据普通搜索引擎(Google、Baidu 等)和学术搜索引擎 Google Scholar 的不同,CAS OpenIR 分别设计了 Meta 映射关系(见表 1)。目前 CAS OpenIR 记录页面中 Meta 值已基本覆盖了论文题录中常见字段。

表 1 CAS OpenIR 映射关系公式

标准映射关系(适用于 Google、Baidu 等搜索引擎)	<dspace-schema>.<element>[.<qualifier>]=<output-schema>.<output-element>[,<scheme>]
针对 Google Scholar 的 Meta 映射关系	<google.citation >_<element>[_<qualifier>]=<dspace-schema>.<element>[.<qualifier>]

Google Scholar 有着自己的一套 Meta 定义规范,因此机构知识库中 Meta 能够被 Google Scholar 索引收录,就需要转换为标准的 Google Scholar Meta。实例对应如下:

google.citation_title = dc.title | dc.title.alternative

google.citation_authors = dc.contributor.author

google.citation_date = dc.date.copyright | dc.date.issued | dc.date.available | dc.date.accessioned

google.citation_keywords = dc.subject.keyword

google.citation_type = dc.type

google.citation_abstract_html_url = $handle

google.citation_pdf_url = $simple-pdf

其中等号左侧为 google 搜索时用到的 Meta 值,可以根据系统中 DC 字段实际情况进行增加、删除、扩展。其中 $handle 变量从具体条目的 handle URL 获取值;$simple-pdf 变量获取完整的全文 URL,前提是条目必须有唯一全文并且为 pdf 格式。

以中科院国科图机构知识库[2]中条目 http://ir.las.ac.cn/handle/12502/4703 为例,系统自动在<head>区生成的 Meta 值对最终结果如下:

<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />

<meta name="DC.subject" content="图书情报理论研究与信息社会学" />

<meta name="DC.subject" content="信息组织与服务" />

<meta name="DC.subject" content="信息技术" />

<meta name="DC.subject" content="图书馆管理" />

<meta name="DC.subject" content="实体图书馆" />

<meta name="DC.subject" content="情报研究" />

<meta name="DC.title" content="研究图书馆 2020:嵌入式协作化知识实验室? " />

<meta name="DC.creator" content="张晓林" />

<meta name="DCTERMS.issued" content="2012-01" scheme="DCTERMS.W3CDTF" />

<meta name="DCTERMS.dateAccepted" content="2012-01-18T05:45:29Z" scheme="DCTERMS.W3CDTF" />

<meta name="DCTERMS.available" content="2012-01-18" scheme="DCTERMS.W3CDTF" />

<meta name="DC.identifier" content="http://ir.las.ac.cn/handle/12502/4703" scheme="DCTERMS.URI" />

<meta name="DCTERMS.bibliographicCitation" content="张晓林.研究图书馆 2020:嵌入式协作化知识实验室? .中国图书馆学报,2012,38(197):11-18" />

<meta name="DC.type" content="期刊论文" />

<meta name="citation_title" content="研究图书馆 2020:嵌入式协作化知识实验室? " />

<meta name="citation_pdf_url" content="http://ir.las.ac.cn/bitstream/12502/4703/1/111221-%e7%a0%94%e7%a9%b6%e5%9b%be%e4%b9%a6%e9%a6%862020-%e4%b8%ad%e5%9b%bd%e5%9b%be%e4%b9%a6%e9%a6%

86%e5%ad%a6%e6%8a%a5%ef%bc%8d201201.pdf" />

<meta name = "citation_abstract_html_url" content = "http://ir.las.ac.cn/handle/12502/4703" />

<meta name="citation_issue" content="197" />

<meta name="citation_date" content="2012-01" />

<meta name="citation_authors" content="张晓林" />

<meta name="citation_keywords" content=" 研究图书馆；数字图书馆；知识服务；知识管理；知识实验室" />

<meta name="citation_volume" content="38" />

4.3　构建 IR 动态网站地图

由于目前大部分搜索引擎只跟踪网站内有限数量的链接，例如 Google 并不会主动抓取网站的所有页面，尤其是网址里带有“?”的动态链接。因此，当网站较大时，例如 IR 会随着学术产出的逐年不断增长而页面快速增多，就必须有有效的策略来保证 IR 中每一条记录目页面都可以被搜索引擎收录。目前来看，通过生成和提供网站地图(sitemap)已成为一种相对可靠的策略和方法。

在 Google 官方指南中可看到，网站生成 SiteMap 文件将有利于搜索引擎机器人的索引，会大大提高索引网站内容的效率和准确度。SiteMap 主要有以下作用[11]：

＊为搜索引擎机器人提供可以浏览整个网站的链接；

＊为搜索引擎机器人提供一些链接，指向动态页面或者采用其他方法比较难以到达的页面；

＊作为一种潜在的着陆页面，可以为搜索流量进行优化；

＊如果访问者试图访问网站所在域内并不存在的 URL，那么这个访问者就会被转到“无法找到文件”的错误页面，而网站地图可以作为该页面的“准”内容。

目前 sitemap 地图在网站应用中越来越受重视，但是人工制作 sitemap 地图的难度随着网站网页数目的增多也变得越来越困难。因此，CAS OpenIR 系统中设计增加了自动生成和发布 SiteMap 的功能，系统自动索引内部所有记录页面生成索引文件(SiteMap)，不限制数量和深度。CAS　OpenIR 中 SiteMap 流程图(见图 3)如下：

①SiteMap 模块触发索引机制后生成 sitemaps 文件，一般会根据系统内页面链接的数量生成 1 个主索引文件(索引文件的索引文件)和 10~50 个二级索引文件。

②在创建好站点地图后，需要主动将其提交给搜索引擎，节省收录时间。使用 Google Webmaster Tools 工具提交 sitemaps 后，会生成相应报表（见图 4），显示已提交 URLs 数量、被收录 URLs 数量、被搜索信息、URL 错误信息等。

③使用 rebots.txt 文件中添加 sitemap 地址的来自动提交 sitemap。

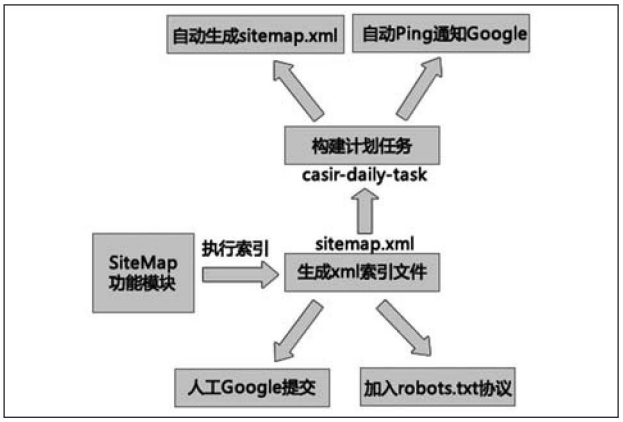

图 3　CAS OpenIR 中 SiteMap 流程图

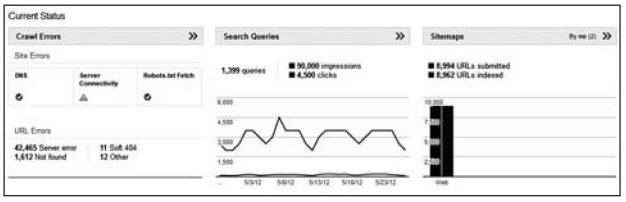

图 4　Google Webmaster Tools 中 sitemaps 反馈统计图

④编写批处理脚本文件，以触发时间节点的定时执行 sitemap 索引任务。

⑤通过 Ping 请求向 google 提示。Ping 是基于 XML_RPC 标准协议的更新通告服务，用于内容更新快速通知给搜索引擎，以便搜索引擎及时进行抓取和更新。因此当 IR 中内容发生了改变，会生成不同的 sitemap 索引文件，此时需要通过 Ping 请求通知搜索引擎进行重新收录。

4.4　其他 ASEO 策略和方法

在 CAS OpenIR 支持 ASEO 优化过程中，同时采用了以下多种辅助性的策略和方法来进一步丰富和完善其整体 ASEO 方法框架。

(1) 优化配置 Robots 协议文件。通过界定 Robots 搜索引擎收录规则，告知 Google Scholar 机器人哪些页面可以收录，哪些页面不能收录。同时使用 Robots 协议告知搜索引擎有关站点地图 SiteMap 的信息。在 robots.txt 文件中包含 SiteMap 链接的好处是，开发人员不用到搜索引擎的站点管理员页面去提交自己的 sitemap 文件，搜索引擎的机器人会主动抓取 robots.txt，读取其中的 sitemap 路径，接着进行相关页面的抓取和索引。

(2) 动态 URL 优化。IR 的一些页面使用动态的

URL,往往附带有很多参数,并比较长,会不利于搜索引擎收录和提升排名。因此,这对这一问题,主要通过 URL 重写的方法[12]进行了优化调整,以获得伪静态和简洁友好的 URL 网址。如 IR 动态生成的 URL 地址 http://[IR 域名]/profile?action =eperson -profile&unique_id =0 -000343,通过重写和优化后将成为 http://[IR 域名]/ unique_id=0-000343。

(3) 英文场景 SEO 优化。解决英文场景下的 Google Scholar 对 IR 的收录和索引。CAS OpenIR 目前通过定制中英文字符集,提供中文、英文两种字符描述,在英文环境下,栏目分类、导航、指引文字均为英文描述,并且页面 Meta 标签组包含有英文题名、英文关键词、英文摘要,可以被搜索引擎英文状态所搜索收录。

5 ASEO 实践效果

CAS OpenIR 在 ASEO 前,学术内容在 Google、Google Scholar 中被索引的情况较不理想。本文选择未进行 ASEO 功能优化的中科院遥感所 IR(http://ir.irsa.ac.cn)为例,其中内容 2906 条,Google Scholar 中被索引率为 0(见图 5)。

图 5 IRSA 在 Google Scholar 中搜索结果图

经过 ASEO 技术全面改进后,在 Google Scholar 中,笔者以中科院国家科学图书馆机构知识库(http://ir.las.ac.cn)为例进行搜索,显示“About 516 results (0.14 seconds)”。意即这 516 篇论文不仅题录信息,其全文也纳入了 Google Scholar 的全文检索。

6 结语

增强 IR 内容发现和利用影响非朝夕工作,是一项系统工程,需要大量的积累和尝试。其中 ASEO 过程已不仅是技术,而是一种思想,一种策略,许多技巧的组合。通过 ASEO 策略可以将机构知识库收录入学术搜索引擎中,在科研人员和学生使用搜索引擎科研过程中,无缝推介和曝光 IR 内容。下一步,我们会继续提高 CAS OpenIR 学术内容在搜索引擎中的索引收录率,使 IR 和其中的论文得以充分可见,积极提高 IR 内容发现和利用影响力。本文中基于 SEO 策略的增强知识内容发现和利用影响的实践过程,对其他数字图书馆服务系统也有着积极的借鉴作用和意义。

参考文献:

[1]OpenDOAR chart[EB/OL].[2012-06-18].http://opendoar.org/find.php?format=charts.

[2]Brody, T. and Harnad,S. Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals [J/OL].[2012-07-10].http://eprints.ecs.soton.ac.uk/10207/.

[3]Arlitsch,K.and O'Brien P.Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar [J].Library Hi Tech, 2012, 30(1):60-81.

[4]苏悦,张文德.Google Scholar 与现代图书馆[J].情报探索,2007,(11):10-12.

[5]祝忠明.中国科学院机构知识库建设软件[R].Post-Conference of Berlin 8 Open Access Conference,2010.

[6]Search Engine Optimization(SEO)[EB/OL].[2012-05-25].http://zh.wikipedia.org/wiki/SEO.

[7]Dirk Lewandowski,Philipp Mayr.Exploring the Academic Invisible Web[J].Library Hi Tech. 2006,24(4):529 - 539.

[8]Google.Inclusion Guidelines for Webmasters [EB/OL].[2012 -06 -18].http://scholar.google.com/intl/en/scholar/inclusion.html.

[9]Page,L.,Brin,S.,Motwani,R.andet al.The PageRank Citation Ranking: Bringing Order to the Web [EB/OL].[2012 -06 -18].http://citeseerxist psu.edu/viewdoc/summary?doi=10.1.1.31.1768.

[10]Dublin Core Collection Description Application Profile [EB/OL].[2012-05-10]. http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/.

[11]Sitemap[EB/OL].[2012-05-18].http://zh.wikipedia.org/wiki/Sitemap.

[12]Rewrite engine[EB/OL].[2012-02-25].http://en.wikpedia.org/wiki/Mod_rewrite.

作者简介:卢利农(1985-),男,中科院国家科学图书馆兰州分馆馆员 祝忠明(1968-),男,中科院国家科学图书馆兰州分馆研究员;张旺强(1985-),男,中科院国家科学图书馆兰州分馆馆员 刘巍(1980-),男,中科院国家科学图书馆兰州分馆馆员;姚晓娜(1985-),女,中科院国家科学图书馆兰州分馆馆员。