现代情报
2007年9月 第9期
September.2007 No. 9

开放式引文作为情报分析源的可行性分析

崔景昌[1,2] 刘德洪[1]
（1. 中国科学院国家科学图书馆，北京 100080；2. 中国科学院研究生院，北京 100049）

〔摘 要〕 本文首先引入了开放式引文的概念与范畴，然后从必要性与充分性两方面论证了开放式引文作为情报分析源是可行的。同时也指出了对开放式引文进行分析时，要注意数据适用范围，方法的局限等问题，以期能得到有科学意义的分析结果。

〔关键词〕 开放式引文；开放获取；引文分析；情报分析

〔Abstract〕 The paper introduces the concept of open citation firstly, and then demonstrates the feasibility of open citation as intelligence source. Meanwhile, the paper points out that while analyzing open citation, we must know the source data's function scope and every method's drawbacks, so that we can get scientific analysis results.

〔Key words〕 open citation; open access; citation analysis; intelligence analysis
〔中图分类号〕G352 〔文献标识码〕A 〔文章编号〕1008－0821（2007）09－0028－04

理论探索

随着开放获取运动的稳步推进，越来越多的机构、组织和国家纷纷加入这种新的出版形式，使得传统的科学交流模式也在发生着重大变革。文献的开放极大的提高了科学信息交流的效率和使用频率，扩展了情报源范畴。那么相应的对于文献、作者或机构等的评价问题，自然应该考虑纳入这种新形式的情报源，使其成为情报分析源。本文在阐释开放式引文的基础上，对其作为情报分析源的可行性给予一定的分析。

1 开放式引文

在因特网环境下，引文已经成为了一种规范化的链接标识，它除了同时具有原来的引文功能之外，又具有了链接功能，它所链接的是文献或者说是知识，也有学者称其为知识链接。在引文成为链接的前提下，我们希望通过引文能够在不同的相关文献间进行自由的跳转，这也正是CrossRef在做的事情。但由于不同的文献分布于不同的数据库，分属于不同的出版商，访问权限成为开放链接的主要障碍，也就还无法达到开放的目的。但对于开放获取文献来说，不存在这样的问题，OpCit（The Open Citation Project）项目即将开放式引文（Open Citation）描述为开放获取文库间引文与全文的互联互通，开放式引文也就是指开放获取文献。笔者在这里将开放式引文的概念更进一步的扩展。我们知道，传统的学术论文撰写从最初的文献获取到论文刊出的整个过程都是封闭的，引文是在较封闭的环境下被作者引用以及被编辑加工，所以在正式刊出之前它们是无法被作者和编辑以外的其它人发现和引用，从而也无法被学术界的新进展所影响。如果打破这种封闭的环境，从学术论文录用开始甚至从撰写开始，即将引文置于某种共享的环境下，其它所有有兴趣的人都可以开放的获取和使用，这是一种新形式的开放式引文。那么我们把涵盖开放获取的，以及在共享环境下使用的引文称为开放式引文。

开放式引文包含三层含义，一是作者的开放式行为，如将自己论文投诸相应期刊外，还把论文挂靠于某种可开放获取的环境，二是各开放获取库通过引文进行互联，三是读者的开放式使用，如引文可像书签一样共享，在一个大的网络社区或团体内开放，如CiteULike，以及《自然》杂志社的引文共享工具克罗提（Connotea）等。

为了对开放式引文作为情报分析源的可行性进行分析，我们把开放式引文分为如下几类：一是开放获取期刊中的引文，一是开放获取数据库或机构知识库中的引文。当然这二者之间有相当一部分的文献是交叉的。本文中的情报分析限于以引文信息为基础的分析，如评价文献、作者或机构，挖掘作者之间关系，发现研究热点等。本文的主旨是论证开放式引文应不应该以及能不能进行这样的利用，也就是从必要性与充分性两个方面，对开放式引文作为情报分析源的可行性进行探讨。

2 开放式引文作为情报分析源的必要性分析

传统的学术出版仍在科学交流中发挥着重要作用，但面临的问题也很多，开放获取最初就是为了解决“学术期刊危机”和“获取危机”问题而发起的，从“布达佩斯倡议”到“柏林宣言”，从arXiv.org到Google Scholar，其影响力越来越大，但到目前为止也还有很多质疑的声音，在利用引文进行分析和评价的时候，往往都没有考虑纳入开放式引文。那么到底有没有必要将开放式引文囊括进情报分析源呢，我们先来看看如下一些统计和事实。

2.1 开放式引文的规模不断壮大

1992年只有5种期刊提供所出版文章的开放获取，但近10多年来开放获取期刊的规模正在不断发展壮大。目前，据DOAJ的统计，它收录了2 618种同行评议的开放获取期刊，130 015篇文章，而J-Gate的收录数目达到了3 835种。全世界大约出版25 000种科技与学术期刊，开放获取

收稿日期：2007—06—07
作者简介：崔景昌（1982—），男，中国科学院国家科学图书馆硕士研究生。
刘德洪（1962—），男，中国科学院国家科学图书馆研究员，主编、参编图书和文集2部，发表和撰写研究报告及学术论文20多篇。

期刊在其中占的份额已经超过了15%，这已经是一个非常可观数字了，但再来看一下增长率。由于笔者没有记录以前的数据，所以这里借用一下北京大学博士李武和中科院初景利教授分别在2005年7月以及2006年9月的DOAJ数据，再加上笔者的数据，时间跨度为一年半，增长率接近50%（见图1）。由此可见开放获取期刊已经达到了一定的规模，并且还在以一个较快的速度在增长。

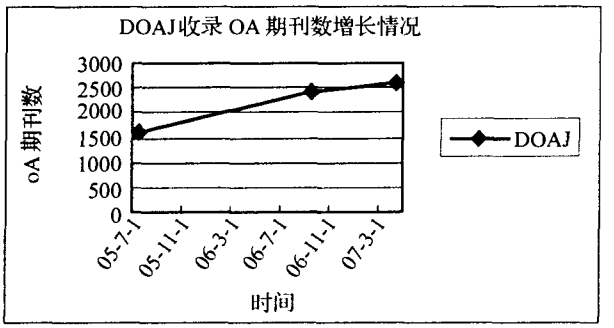

图1 DOAJ收录开放获取期刊增长情况统计

从1991年arXiv.org的建立到现在，开放获取机构知识库也已经达到了相当的规模。据ROAR（Registry of Open Access Repositories）的统计，截止2007年4月1日，在该机构登记的机构知识库已达862个，这其中包括一些电子期刊和出版物。具体增长如图2所示，单纯的机构知识库接近700个，文章数接近600 000篇。DOAR（Directory of Open Access Repositories）统计数字为853个。由此可见大部分的开放获取文章都存于机构知识库中，随着越来越多的国家和机构对机构知识库建设的重视和政策上的倾斜，这部分的开放式引文上升空间还是巨大的。

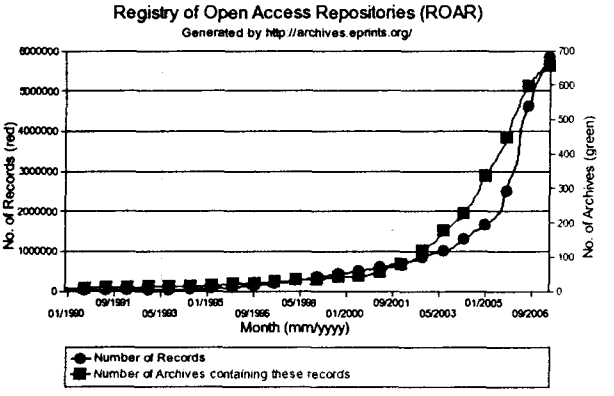

图2 ROAR收录的机构知识库增长情况统计

开放获取文献散布于各种开放获取期刊、机构知识库、预印本系统、学科资源库以及其它地方，这种状况促使了统一标准的建立。开放存档计划元数据收割协议（Open Archives Initiative Protocol for Metadata Harvesting，OAI－PMH），提供对可以用于建立开放获取数据库的工具的链接，实现检索系统（服务提供者）从各种资源库和知识库中检索开放获取文献的元数据，并对这些数据加以集成以便一次提问就能检索到所有的结果。有了统一的接口，开放获取文献的社会化开放化使用就变得更为便捷。出现了很多开放式引文使用工具，如搜索工具Citebase、Citeseer、Google Scholar等，社会化书签使用工具CiteULike、Connotea等。这些工具更是极大的促进了开放式引文的使用，从文献搜索、获取到撰写论文时文献信息的插入（上述工具都可以把引文元数据传输到参考文献管理软件中，因此写作时参考文献的插入可以实现自动化），基本上可以做到很少量的人工参与，很少的时间耗费，无形的扩大了开放式引文使用的规模。

2.2 开放式引文的影响力不断提高

早期的开放获取期刊和机构知识库只集中于少数的国家和少数的几个学科，到现在，这种地域性和学科性有了明显的突破。以DOAJ为例，它收录的开放获取期刊目前覆盖的学科领域已经涉及农业和食品科学、艺术和建筑、生物和生命科学、语言和文学等17个学科与专题领域，突破了自然科学的局限。开放式引文所覆盖的地域范围也开始遍布全球各大地区，很多发展中国家也开始注重发展开放获取期刊和机构知识库。图3为DOAR收录的世界范围内开放获取机构知识库按洲的分布情况，可以看到南美和亚洲已经占到了相当的比例。

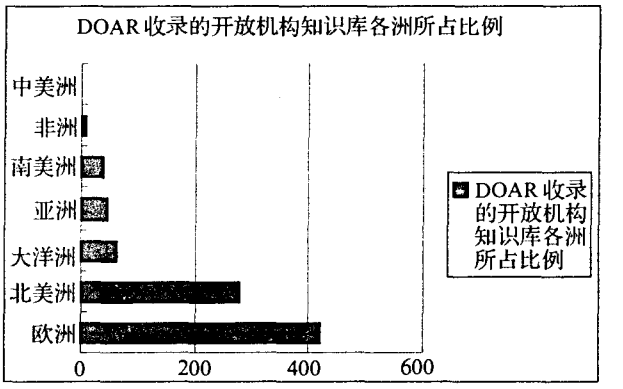

图3 DOAR收录的开放获取机构知识库各洲所占比例情况

文摘和索引是信息服务机构对学术论文进行评价的重要手段，若期刊被权威的文摘索引数据库收录，则意味着期刊论文被检索和发现的概率大大提高，同时也就意味着论文被引用的可能性更大。开放获取期刊发展的一个重要事实便是开始得到传统的文摘索引服务商的认可并成为它们收录的对象。据ISI期刊引用报告显示，截至2005年5月，被SCI收录的开放获取期刊已经有270种，目前已不限于这个数字。ISI收录开放获取期刊本身就证明被收录的部分已经达到了所限定的阈值，但由于SCI的收录特性，也决定了很多高质量的开放获取期刊没被收录进去。Stevan Harnad和Tim Brody的研究表明开放获取可极大的提高文章的被引频次，比如，在电子工程学科中，发表于同一种期刊中的OA论文的平均被引次数为2.35，非OA论文的平均次数为1.5。很多对各学科的统计研究都得到了类似的结论。

机构知识库的情况同样如此，尽管其中存储的文章很多都没有经过同行评议和编审，但其学术影响也同样是很高的，不乏优秀之作。2006年素有数学诺贝尔奖之称的菲尔茨奖得主格里高利·佩雷尔曼，就是将其解决了“庞加莱猜想”的3篇重头论文简单的提交到了互联网上的一个数学文献库里，但他的思想同样得到了国际数学界的充分重视和认可。这不可不谓是对传统学术出版的一个成功挑战。开放式的获取与使用提高了学术交流的效率，也使得人们

更关注研究问题本身，而不是资料的获取。

开放式引文工具在无形中扩大了开放式引文的规模之外，也切实的提高了开放式引文的影响力。通过字段的简单限定，就可通过 Scirus、Google Scholar 等工具迅速定位到目标文献，并且这些工具会自动的进行扩展搜索，为用户提供相关文献。通过很简单的操作，Connotea、CiteULike 就可以把用户在网上找到的文章的元数据抽取出来，形成用户个人的资料库。通过简单的转换，引文的元数据就可导入到 EndeNote 等参考文献管理软件中，随时随地的查阅和撰写论文插入引文元数据。文献的质量，再加上使用上的便利，开放式引文在学术交流中已开始发挥十分重要的影响。

随着开放式引文的规模不断扩大，影响力不断提高，我们在利用引文信息进行分析评价工作的时候，不能不考虑吸收开放式引文，以更科学更客观的进行分析评价。因此开放式引文作为情报分析源是十分必要的，不把开放式引文作为有机组成部分的引文分析是不完整的。

3 开放式引文作为情报分析源的充分性分析

既然开放式引文作为情报分析源是必要的，如何对其进行分析自然就成为我们要考虑的问题，如果有切实可行的方法，能得到有价值的分析结果，那么我们进行分析就是有据可依有章可循的。正如 Jeffrey M.Perkel 在《引文分析的未来》中指出，如何追踪以非传统形式出版的文章的影响，是引文分析面临的挑战之一。该部分对已经出现的一些开放式引文分析方法、可用来进行分析的以及具有分析潜质的数据进行一定的归纳。

3.1 利用引文索引工具

引文索引商现在已经十分重视开放获取期刊的收录，ISI 正逐步加大这方面的力度，新力军 SCOPUS 从建立之初就十分重视非常规渠道出版物的收录，在其收录的15 000多种期刊中，开放获取期刊有 500 多种，会议文献有 700 多种。这部分的数据将能够支持我们以传统的方法进行引文分析。在近年来证明开放获取文献比传统文献具有引用优势的研究中，用的基本上都是基于这种成熟引文索引工具的引文分析。如 Lawrence 在对计算机科学方面开放获取与非开放获取的会议论文平均引用次数的比较，Antelman 在政治、哲学、电子工程以及数学四个领域开放与非开放的对比研究等等。

权威的引文索引工具数据较为可靠，已经成为了科研评价的权威数据来源。对开放式引文的评价，也自然离不开这些工具的支持，但其并不足以揭示开放式引文的全貌与使用特性。

3.2 利用大型的学术搜索引擎

商业搜索引擎，如 Google 和 Yahoo 已涉足学术搜索领域，开放获取文献成为他们重要的索引对象，前者使用的是 OCLC 开发的 OAI 入口，后者则与 OAIster 展开了广泛的合作。在大量的占有文献信息的基础上，搜索引擎也具有了自动统计引文信息的可能。如 Google Scholar 在呈现搜索结果的同时，列出了文章的被引次数，相关的作者等信息，见图 4。示例文献被引用了 145 次，继续点击的话可查看引用文献。在高级搜索里可限定元数据各字段。

All Results
D Knuth
S Papert
J Hennessy
M Garey

Computer-based real-time conferencing systems - group of 3 »
S Sarin, I Greif - **Computer** 1985 - portal.acm.org
... **Computer**-based real-time conferencing systems. Source. **Computer** archive Volume 18 , Issue 10 (October 1985) table of contents. Pages: 33 - 45. ...
Cited by 145 - Related Articles - Web Search - Find in ChinaCat

图 4 Google Scholar 学术搜索结果呈现形式

搜索引擎对文献的标引是机器自动完成的，在现阶段也还存在着不少问题。Peter Jacso 指出 Google Scholar 在很多方面已可以与 SCI 相媲美，但经常会出现一些莫名其妙的错误，如将页码识别为年代，被引次数有时候会很离谱等。如果这些问题都解决了，那么我们就有理由相信以学术搜索引擎的数据为基础的引文分析会是对传统引文分析的一个重要参考。尤其是对于开放式引文来说，更是如此，有的可能在传统的索引工具中并没有收录，搜索引擎提供的数据就显得更加珍贵。一些学者，如 Kayvan Kousha、Mike Thelwall 等开展了一些基于学术搜索引擎数据的利用研究。

另外，根据美国一些物理学家所作的一项统计分析，Google 的 PageRank 算法可以提供一种评价科学文献影响程度的好方法，它甚至有可能替代传统的引用率指标。波士顿大学的 Sidney Redner 和 Pu Chen 以及布鲁克海文国家实验室的 Huafeng Xie 和 Sergei Maslov 建议用 Google 的 PageRank 算法来发掘这样的重要文献。他们在研究中指出，用文章被引用的次数衡量文章的重要性并不总是有效的，比如，它会忽略那些被引用次数相对比较少但是对物理学确有极其重大影响的文章。其中一个例子是 Richard Feynman 和 Murray Gell - Mann 1958 年发表的“费米子相互作用理论(Theory of the Fermi Interaction)”，他们在这篇文章中引入了后来成了“标准模型”中关于弱相互作用的理论，利用新方法就能发现这篇文章的重要性。由此启发，在引文也成为链接的前提下，链接分析的方法也可应用于引文分析中。

3.3 利用小型的开放式引文搜索引擎

这类工具中的典型代表是 Citebase 和 Citeseer，不像大型综合性搜索引擎存在统计粗糙的问题，它们在其自身的收录范围内，提供了精细的统计数据，包括点击次数、下载次数以及被引次数，还包括同被引、引文耦合等。这些数据为我们计量其收录范围内的文献提供了丰富的数据支持，并且以直观易懂的可视化界面呈现，图 5 为 Citebase 对一篇文献的引文统计信息。上面一条曲线表示引用该文献的文章数增长情况，对应左边坐标，下面曲线为点击查看情况，对应右边坐标，并详细的给出了时间分布，对引文的情况一目了然，如引文的峰值年代等等。到目前为止该文章被下载了 73 次，有 8 篇文章将其作为引文，2000 - 2001 年文章被引用的次数最多。

这些小型的开放式引文搜索引擎，不但提供了引文信息，还统计了文献的点击下载情况，这部分数据可以反映文献被关注和使用的状况，Tim Brody 等学者已经就下载数据与引用数据之间的关系进行了深入研究，以期能通过下载情况对引用情况进行预测。

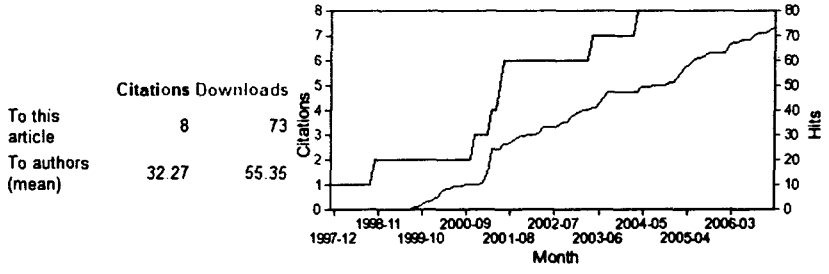

图5 Citebase对一篇文章进行的引文统计图(截于2007年3月16日)

3.4 利用开放数据库的用户使用数据

数据库用户使用数据完整而详细地记载了数据库中各期刊和文献的使用情况，是对数据库使用情况的客观反映。对数据库使用数据进行分析，是一种研究用户访问行为和数据库利用情况的有效方法，因而数据库的用户使用数据一直是数据挖掘的重点研究对象。数据来源主要是服务器日志和数据库的自动统计数据。大部分开放数据库将统计数据实时公布，供人们查阅，如arXiv，服务器日志数据就只能是拥有数据库的机构才能看到并利用了。数据库用户使用数据的分析同样能够追踪文献前期的使用情况，并对后期的引用作出适当预测，也是引文分析的有益补充。Philipp Mayr即基于用户使用数据提出了使用影响（usage impact）和使用偏好（user preference）指标，作为引文影响的补充。

3.5 利用开放使用社区中的数据

开放式引文使用社区，像CiteULike和Connotea，以书签化的方式对引文进行收录，以用户自由赋标签（tag）的自由分类法（folksonomy）的形式对引文进行标引，以共享的理念进行使用。可以说这种开放式引文使用社区就是一个大的引文索引工具，由广大的文献使用者进行索引，每个用户由自己的需求拉动，大量的用户行为进行积累，便使得这个工具具有了社会性，以最直接最快速的方式反映了用户的关注内容。开放式引文使用社区中的数据也可用于追踪开放式引文的关注和使用情况，作为开放式引文分析的参考。目前也有学者在这方面展开了尝试。

由此我们可以看到，无论是开放获取期刊中的文献，还是开放获取机构知识库或集成的开放式引文搜索工具中的文献，都有一定的分析方法对其进行引文分析，为相关决策和评价提供情报支持。因此将开放式引文作为情报分析源是可行的，能够得到具有科学意义的分析结果。但同时也要注意来源数据的适用范围，每种分析方法的局限性等问题，应尽量采用多种方法综合相互补充与修改，以免造成数据滥用。

4 小 结

随着各国对开放获取的重视，开放获取的规模和影响不断扩大，各种针对开放式引文的分析方法也不断涌现，开放式引文成为情报分析源既是必要的也是充分的。积极的探索出适合开放式引文的分析和评价方法是目前图书情报界需重点关注的问题之一，从而能够更科学的进行引文分析，为科学决策提供更有参考意义的情报，继而更好的促进科学交流模式的发展。

参 考 文 献

[1] 毛军. 社会化引文网络和科学范式的重建 [J]. 图书情报工作，2006，(9)：31－35.

[2] 贺德方. 知识链接发展的历史、未来和行动 [J]. 现代图书情报技术，2005，(3)：11－15.

[3] Open Citation（OPCIT）Project [EB]. http://opcit.eprints.org/opcitabout.shtml [2007－03－20].

[4] The Directory of Open Access & Hybrid Journals [EB]. http://www.doaj.org [2007－03－30].

[5] J-Gate [EB]. http://www.j－gate.informindia.co.in/ [2007－03－30].

[6] 李武，杨屹. 开放存取期刊出版的发展现状及其影响分析 [J]. 图书情报工作，2006，(2)：25－30.

[7] 初景利. 开放获取的发展与推动因素 [J]. 图书馆论坛，2006，(6)：238－242.

[8] ROAR（Registry of Open Access Repositories）[EB]. http://roar.eprints.org/index.php [2007－04－01].

[9] Stevan Harnad, Tim Brody, Comparing the Impact of Open Access（OA）vs. Non－OA Articles in the Same Journals [J]. D－Lib Magazine, 2004, (6). http://www.dlib.org/dlib/june04/harnad/06harnad.html

[10] Jeffrey M Perkel, The Future of Citation analysis [J]. The Scientist, 2005, 19 (20): 24.

[11] Steve Lawrence, Free Online Availability Substantially Increases a Paper's Impact [J]. Nature, 2001, (5): 521.

[12] Kristin Antelman, Do Open Access Articles Have a Greater research Impact? [J]. College & Research Libraries News, 65 (5): 372－382.

[13] Peter Jacsó, As we may search. Current Science [J]. 2005, (11): 1538－1547.

[14] Kayvan Kousha, Mike Thelwall. Google Scholar Citations and Google Web/URL Citations: A Multi－Discipline Exploratory Analysis [EB]. http://eprints.rclis.org/archive/00006416/ [2007－04－05].

[15] Belle Dumé, Google unearths physics gems [EB]. http://physicsweb.org/articles/news/10/4/10/1 [2007－04－05].

[16] Tim Brody, Stevan Harnad. Earlier Web Usage Statistics as Predictors of Later Citation Impact [EB]. http://www.ecs.soton.ac.uk/~harnad/Temp/timcorr.doc [2007－04－05]

[17] Philipp Mayr, Constructing experimental indicators for Open Access documents [EB]. http://arxiv.org/abs/cs.DL/0610056 [2007－04－05].