

借助文章级附加信息提升数字资源系统用户服务能力

□ 王欣 / 高等教育出版社 北京 100029

摘要: 文章级附加信息帮助用户评估单篇文章价值并发现相关文献网络。文章重点介绍PLoS文章级计量信息的六个维度, 相关引证文章、相关主题文章、相关作者与机构文章的类型与应用实例, 增强系统易用性并与写作环境相融合的举措, 提出数字资源系统应利用第三方专业服务整合优质资源, 借助文章级附加信息提高搜索效率, 重视Web2.0和Web3.0技术应用。

关键词: PLoS, 文章级计量信息, 相关文献, 数字资源系统, 用户服务

DOI: 10.3772/j.issn.1673-2286.2011.10.003

1 导言

近年来, 越来越多的数字资源系统, 无论是来自于商业出版社、Open Access机构、图书馆、公益性研究项目还是第三方中介服务, 在保障资源数量与质量的前提下, 都着力于丰富文章级附加信息, 以期提升用户服务能力, 增强更广泛用户对系统的依赖感和忠诚度。

这些附加信息可以帮助用户评估单篇文章价值并形成文献关系网络。具体地, 当用户通过数字资源系统知识组织体系及搜索服务获得相关文章列表并且到达一篇指定文章后, 除了阅读详细的元数据及全文信息, 用户可以通过文献计量与网络计量多维指标体系辅助评价文章价值与学术影响力, 进而通过相关文章网络理清科研脉络、推动科学发现。这些文章评价与关系网络数据若被知识组织与搜索服务所利用, 则可在一定程度上提高搜索效率。所有的文章内容及其附加信息尽可能以方便用户使用的方式提供, 并尽可能与用户科研及写作环境相融合(图1)。

目前还没有哪一个数字资源系统完整提供如上全部文章级附加信息与服务, 但是有不同的系统在其中的某些方面作出积极尝试。

2 文章级计量信息

国内外常用的评价体系大多基于该文章所发表的期刊进行, 缺乏对文章本身进行评价的成熟工具与方法。近几年的研究中, A. Schubert^[1]和L. Bornmann^[2]分别在不同的文章中讨论h指数在单篇文章评价方面的应用; A. Thor基于Google Scholar数据开发的自动计算并可视化单篇文章h指数及其他相关指标的网络应用系统将单篇文章h指数评价从理论研究推向系统验证阶段, 但尚未有成熟应用^[3]。目前真正在数字资源系统中提供服务的只有美国PLoS(Public Library of Science, 公共科学图书馆)制定的文章级计量信息^[4](Article-Level Metrics Information), 它从6个维度全面考量学术文章的内在价值, 具有一定的指导意义。

PLoS认为研究文章的学术水平不能完全由文章所属期刊的影响力来决定, 主要还是依赖于文章自身价值和贡献度。它于2009年3月启动项目, 为其所出版期刊的每篇文章提供文章级计量信息, 具体包括文章被引数据、使用统计数据、社会化书签数据、评论与批注数据、与该文章有关的博文信息以及来自于用户的文章评分等, 这些计量信息连同文章本身都是开放获取的^[5]。

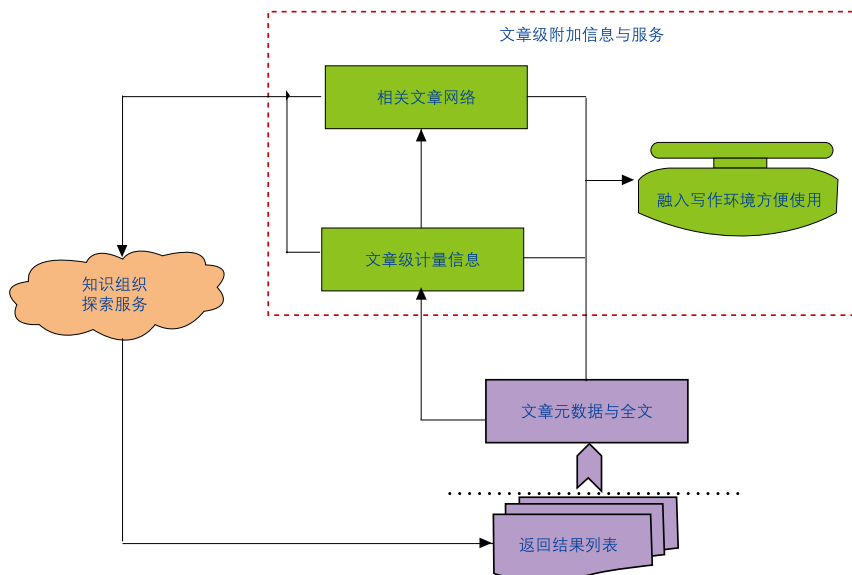


图1 数字资源系统的文章级附加信息

Total Article Views: **24625** from Jun 29, 2007 (publication date) - Aug 6, 2011*

Breakdown by View Type

HTML Page Views: **21976**

PDF Downloads: **2478**

XML Downloads: **171**

Cumulative Views from Jun 29, 2007 (publication date) - Aug 6, 2011*

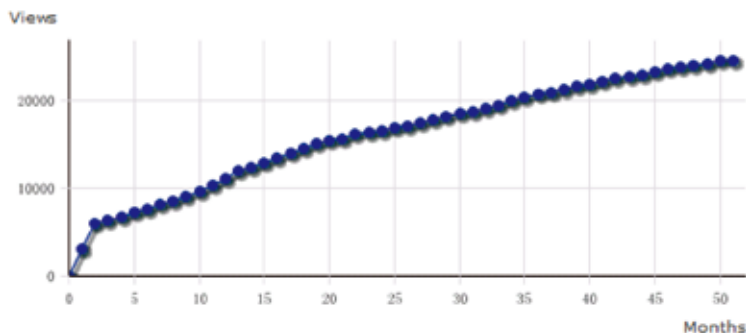


图2 PLoS Genetics期刊某文章使用统计数据示例

2.1 文章被引数据

作为文章级计量信息的重要组成部分，PLoS为每篇文章提供被引数据。这些数据揭示的并不是文章在PLoS系统自身数字资源范围内的被引情况，而是分别取自于被学术界普遍认可的提供引文服务的第三方系统，包括Web of Science、Scopus、PubMed Central及CrossRef。具体地，PLoS利用各个第三方系统提供的应用程序接口自动抓取指定文章的引用数据。来自各个第三方系统的引用数据在PLoS系统中分别显示为一个带有超链接的数字，表明该文章在指定系统内的总被引频次，点击链接后可直接到达第三

方系统获取相应的施引文献信息。需要指出，尽管如上第三方系统收录文献标准各不相同，但大多覆盖所有学科领域，且文献量巨大。这些第三方系统的统计源数据会有相当比例的重复，这也直接导致PLoS从不同第三方系统获得的被引数据之间会有重复。

近年来，Google Scholar作为功能强大的学术文献搜索引擎，在揭示学术文章被引情况方面备受关注。尽管Google Scholar尚未发布应用程序接口，无法自动获得指定文章的被引频次，但PLoS提供了到Google Scholar获取被引信息的链接，点击链接后将直接传递指定文章参数进行Google Scholar搜索。

2.2 文章使用统计数据

PLoS为每篇文章提供3种格式的全文服务，其中默认格式HTML文件更适合在线阅读和检索，PDF文件在文章存档与打印方面有明显优势，此外还特别提供了用于生成各种格式文件的XML原始数据。在文章使用统计方面，系统遵循COUNTER 3标准，为每种文件格式单独计算使用次数（见图2），使用统计数据更新时延为48小时。

学术文章的使用量与文章发表时间长短及文章所属学科领域关系密切，单纯比较不同文章的使用次数并没有实际意义。为了帮助用户更好地解读使用统计数据，PLoS还专门提供了系统内各个期刊在不同年份和不同研究领域之下的文章篇均使用次数作为参考。

2.3 社会化书签数据

PLoS在选择提供文章级计量信

息时,也充分考虑到Web2.0环境下一系列社交网络和社会化媒体工具之于用户评价数字资源方式方法的冲击。社会化书签服务不仅能辅助发现相关文章、实现同领域科研人员聚类,用好社会化书签数据更能在一定程度上揭示指定文章的受欢迎程度。

PLoS选取CiteULike和Connotea这两个学术研究领域的重量级社会化书签服务,为每篇文章分别提供被CiteULike和Connotea注册用户标记书签的次数,点击数字链接后可到达相应的社会化书签服务页面,从中获得文章在何时被哪些用户标记了书签、具体使用了哪些标签来描述文章主题,还可以到达用户主页浏览他所标记的其他书签信息,尤其是与指定文章拥有相同主题标签的那些文章信息。

2.4 评论与批注数据

PLoS系统搭建了文章读者与作者之间的交流平台,允许用户针对所阅读文章进行整体性评价,或者针对部分章节及语句进行批注。而评论与批注条数同样被汇集在文章级计量信息中,用以辅助揭示其受关注程度。点击数字链接后可以追踪到数据创建时间和创建人,并阅读完整的评论与批注信息。

为了维系健康有序的交流环境,PLoS也特别制定有关此类数据的发布与管理规程,所有评论与批注数据都必须在严格遵守规程的前提下非匿名发布,PLoS会对这些留言及回复进行监控。

2.5 与该文章有关的博文信息

尽管PLoS提供了留言板功能,但并不能绑定用户仅在此处对PLoS文章进行评论。用户完全可能在博客和论坛上留下与文章有关的其他信息,而这些信息对于评价文章本身也非常重要。

与收集文章被引数据类似,PLoS选择借助第三方专业博文集成工具来提取这些信息,包括Researchblogging.org、Nature Blogs、Bloglines以及Google Blogs。由于不同集成商本身集成的博客资源会有一定重复,对于分别从各个集成工具获得的与文章相关的博文数据依然存在去重问题。

除依赖第三方平台之外,PLoS在自身系统中还支持trackback功能。

2.6 星级评分

PLoS系统允许用户针对指定文章,从研究深度、可信度及写作风格三个维度进行非匿名评分,得分指标从1颗星至5颗星,用户也可以选择在进行必要的文字说明。

2.7 小结

目前PLoS的文章级计量信息尚未获得广泛应用,像Springer公司的SpringerLink、Elsevier公司的ScienceDirect等大型数字资源系统在文章级评价信息提供方面仍较多地采用其在自身系统内的被引次数或下载量排名。文章被引需要较长周期,通常在文章发表后的6个月内很难获得,这也更突显出文章级计量信息在众多数字资源系统中的缺失。

3 相关文章网络

概括起来,数字资源系统中一篇文章所附加的相关文章信息可以从相关引证关系文章、相关主题文章、相关作者与机构文章三个方面提供。这三个方面信息互为补充,形成完整的文献网络,用户可以从摸清该领域研究发展的脉络。

3.1 相关引证关系文章

相关引证关系文章主要涉及以下内容:

①文章参考文献:即这篇文章在研究与写作过程中参考并重点引用的文献信息,反映作者研究工作的背景和依据。参考文献信息也是文章自身信息不可缺少的部分,在所有的数字资源系统中都会提供,但是系统帮助用户直接获取参考文献线索的能力并不相同。绝大多数数字资源系统会提供参考文献在自身数据库中的链接信息,较成熟的系统会提供参考文献的第三方交叉链接功能,如CrossRef链接、PubMed链接等等,如有OpenURL解析服务还可以发送相应请求进行查询。

②文章引证文献:即在该文正式发表后,将其作为参考文献的那部分文章信息。引证文献是该文研究工作的继续、应用、发展或评价。目前在数字资源系

统中引证关系也主要限定在自身资源范围内,反映文章在自身封闭系统内的被引用情况。较好的做法是提取专业第三方引文服务系统中的被引数据,如Web of Science、Scopus等,这样能使用户更全面客观地掌握文章引证信息。

③文章共引文献:即与该文有相同参考文献的文章信息,表明有共同的研究背景和依据。相较于文章参考文献及引证文献,共引文献与该文的相关度会有所减弱。现有数字资源系统,尤其是出版商全文数据库较少提供此类信息。

④文章同被引文献:即与该文同时被作为参考文献引用的文献,表明与该文共同作为进一步研究的基础。与文章共引文献信息类似,这部分文献也较少被数字资源系统揭示,但是对于用户发现相关文章也很有意义。

⑤二级参考文献:对于特别有价值的研究论文而言,追溯其二级参考文献——即被该文献的参考文献所引用的那部分文献也很有意义,可以帮助用户更进一步追溯文章的研究背景与依据。

⑥二级引证文献:即该文引证文献的引证文献。更进一步反映该文研究工作的继续、发展或评价。

除Web of Science、Scopus等专门提供引文服务的系统外,国内中国知网的中国学术期刊网络出版总库提供了非常全面的相关引证关系文章信息,包括了如上6项内容。不足之处是该系统所有的引证服务全部基于自身封闭系统进行,尽管从收录文献种类及数量上均有优势,但仍会遗漏一些存储于其他资源系统之上的相关引证文献信息。

3.2 相关主题文章

目前绝大多数数字资源都会提供相关文献(Related Articles)链接功能,通常都是提供在自身资源系统内相关主题文章信息。与搜索服务类似,这些相关主题文章服务在呈现方式上大同小异,但后台实现机制却各不相同,这也直接导致主题“相关性”和响应时间差异。

国外早期的二次文献服务系统的相关主题功能实用性很强,多借助专业受控词表进行文献标引。如PubMed数字资源系统使用的MeSH(Medical Subject Headings)词表,Ei Village 2使用的SHE(Subject Headings for Engineering)、EIT(Engineering

Information Thesaurus)、INSPEC叙词表等^[6]。以PubMed相关文献功能为例,其文献之间的相似度通过它们所共有的特征项来衡量。这些特征项包括MeSH词以及从标题和摘要中抽取出来的关键词,同时根据来源位置不同赋予不同权重,如在标题中出现的词按两次计算。每个特征项都有1个局域权值和1个广域权值,其中局域权值取决于特征项在1篇文献中出现的次数和该文献中包含的特征项总数,广域权值取决于整个PubMed数据库中包含有该特征项的文献数目。2篇文章的相似度值,是它们所共同拥有的特征项的局域权值和广域权值乘积之和。一旦可定量出1篇文章与数据库中其他文章彼此间的相关程度,就能够将相似度值大于阈值的文献输出,并按相似度值倒序揭示。PubMed数据库文献间的相似度值计算都是预先处理好的,所以在用户点击相关文献链接时,系统有较好的响应时间^[7]。

国家科技图书文献中心(NSTL)针对其保有的大量二次文献资源,设计并实现了一个实时相关文献推荐子系统。当用户浏览检索结果中的第i篇文献的摘要、关键词等详细信息时,实时计算该文献与检索结果列表中位于该文献后的文献之间的相似度,选择与其相似度最高的k篇文章作为相关文献进行推荐。由于该系统在用户基于关键词的检索结果空间下计算文献之间的相关性,检索结果空间相对于整个文献空间来说非常小,所以系统效率与准确性较高;同时由于整个推荐过程实时计算并进行推荐,可以较好地适应文献数据库动态发展的变化^[8]。

3.3 相关作者与机构文章

Web of Science和Scopus数字资源系统提供了详细的合作作者网络与合作机构网络信息,但是目前尚未发现这些信息被有效集成到一次文献数字资源系统中。对于大多数系统而言,通常会提供该篇文章作者在封闭系统内发表的所有文章信息以及其所在机构发表的所有文章信息。SpringerLink平台上某一文章作者的其他相关文章信息见图3。

4 增强可用性,与写作环境融合

在尽可能提供丰富的文章级附加信息的同时,数



图3 SpringerLink平台上某一文章作者的其他相关文章信息

Export citation

To export the selected citation + link, select your preferred formats and click Export.

Content format: ☒ Citations Only
☐ Citations and Abstracts

Export format: ☒ RIS format (for Reference Manager, ProCite, EndNote)
☐ RefWorks Direct Export ? About Refworks
☐ ASCII format
☐ BibTeX format

图4 Elsevier的ScienceDirect中Export Citation功能

字资源系统也特别重视用户的使用体验，逐步改进系统中各项信息内容的显示方式，争取用户用更少的点击次数获取更多的内容。这一方面Elsevier的ScienceDirect平台做了较多尝试，如提供Show preview、Full size images等功能。

此外，现在绝大多数数字资源系统，尤其是一次文献系统均向用户提供参考文献导出功能（Export Citation），允许将文章信息按篇或整期导入Reference Manager、ProCite、EndNote等文献管理与写作工具中，方便作者在写作环境下直接使用（图4）。

5 数字资源系统丰富文章级附加信息的几点思考

用户对数字资源系统功能与服务的需求是与用户研究环境与行为方式变化同步的。在创新发展的大环境下，数字资源系统必须持续创新，不断满足并拉动用户需求，才有可能形成用户对系统的长久依赖，在文章级附加信息提供方面也是如此。尽管不同类型的数字资源系统在服务提供方面

应各有侧重，但总体原则是能让用户通过尽可能少的访问与点击，获得最相关的那部分文章信息。

5.1 利用第三方专业服务整合优质资源

在文章评价与相关文章信息提供方面，绝大多数数字资源系统都将待选范围局限在自身封闭的系统内。即便是资源收录范围广泛的数字资源系统，也不可能囊括用户需要的所有文献，所以这种信息评价与推荐方式本身就有很大的局限性。

提升用户服务能力并不意味着数字资源系统在任何功能模块及数据上都要亲力亲为、面面俱到，还需要灵活运用各类第三方专业服务。比如对于被引统计数据，完全可以集成Web of Science、Scopus、CrossRef以及Google Scholar的相关数据到自身系统中，这样用户在获取全文的同时，就可以很方便地在同一平台上获得分布于其他不同数字资源系统内的引用统计数据，而不必再分别到其他各个系统中获取。

但是在第三方服务的选择上还是存在隐患的，比如PLoS早期用于集成博客信息的工具Postgenomic已在2010年被自然出版集团关闭，所以PLoS不得不在此后移除文章级计量信息中与之相关的所有数据。我们需要一种机制来稳固用户—数字资源系统—第三方服务的依存关系，形成三方共赢共发展的良性循环。

5.2 借助文章级附加信息提高搜索效率

用户对于数字资源系统的根本要求是“找对文献找准人”，也

即找到与主题相关的最有价值的文献资料以及发掘出自己关注领域内的活跃人物。文章级附加信息在很大程度上都是用于辅助用户判断一篇研究文章的学术价值,并帮助其寻找相关文献以及与之有共同研究兴趣的学术群体。如果文章级附加信息能够被搜索引擎选择利用,丰富检索结果排序特征项,就能在很大程度上提高搜索效率^[9]。

当然,文章级附加信息作为搜索引擎排序依据还需要满足两个基本条件:一是计量维度尽可能被更多数字资源系统广泛使用,比如PLOS的六维度计量信息;二是各维度采用的计量标准尽可能一致,比如统计下载量的COUNTER标准等等。

5.3 关注Web3.0,用好Web2.0

也许Web3.0还在路上,但是Web2.0已经实实在在

地渗透到广泛的日常科研与学术交流环境中。在文章级附加信息提供方面,很多数字资源系统已经注意到对Web2.0技术工具的合理应用,如Nature及其姊妹刊的每篇文章都同时提供与Connotea、Digg、Facebook、Newsvine、Delicious、Twitter等的交互,前面提到的PLOS收集了部分来自社会化书签和博客的与文章有关的数据等。但这只是迈出了第一步。事实上博客、Twitter等Web2.0工具已逐步成为文章发表后的公共讨论平台,如何捕捉、组织和计量所有来自其中的与文章有关的信息;如何聚合和量化一篇文章的所有在线回复以及如何通过文章的所有在线活动来快速计算文章的影响力,这些问题推动了新型交流模式下一个新学科的产生——选择计量学(alt-metrics)^[10]。

参考文献

- [1] SCHUBERT A. Using the h-index for assessing single publications [J]. Scientometrics, 2009, 78(3): 559-565.
- [2] BORNEMANN L, DANIEL H D. The state of h index research: Is the h index the ideal way to measure research performance? [R]. EMBO Reports, 2009, 10(1): 2-6.
- [3] THOR A, BORNEMANN L. The calculation of the single publication h index and related performance measures: A web application based on Google Scholar data [J]. Online Information Review, 2011, 35(2): 291-300.
- [4] NEYLON C, WU S. Article-Level Metrics and Evolution of Scientific Impact [J]. PLoS Biology, 2009, 7(11).
- [5] Article-Level Metrics Information [OL]. [2011-08-10]. <http://www.plosbiology.org/static/almInfo.action>.
- [6] 李琳, 华薇娜. 受控语言在国内网络数据库中的应用与展望[J]. 中国索引, 2010(4).
- [7] 王军辉, 李丹亚, 等. 国内外部分文献数据库检索系统相关文献功能浅析[J]. 中华医学图书情报杂志, 2010, 19(5): 68-71.
- [8] 张志平, 李琳娜. NSTL文献检索系统中相关文献推荐功能的设计与实现[J]. 现代图书情报技术, 2010(7): 110-113.
- [9] 李枫林, 等. 面向用户潜在信息需求的检索结果组织研究[J]. 情报理论与实践, 2011(5): 42-45.
- [10] MANDAVILLIA. Trial by Twitter [J]. [2011-08-10]. Nature, 2011(469): 286-287.

作者简介

王欣 (1978-), 女, 图书馆学博士, 高等教育出版社自然科学学术出版事业部, 高级编辑。E-mail: wangxin@hep.com.cn

Enhancing User Service Capabilities of Digital Resources Systems through Additional Article-Level Information

Wang Xin / Higher Education Press, Beijing, 100029

Abstract: Additional article-level information helps users evaluate the specific publication and find related articles. This paper focuses on the six dimensions of PLoS Article-Level Metrics Information, types and application cases with citing and cited articles, subject related articles and articles written by the same author or from the same institution, emphasizes that all these kind of information should be molten into authoring environment and easily used. Integrating high quality resources from third-party services, improving search efficiency using additional article-level information and attaching importance to web2.0 and web3.0 technologies have been suggested to enhance user service capabilities of digital resources systems.

Keywords: PLoS, Article-level metrics information, Related articles, Digital resources system, User service

(收稿日期: 2011-08-10)