

电子预印本文献被引用的实证分析

曹兰芳 查先进

(武汉大学信息资源研究中心, 湖北武汉 430072)

摘要: 文章利用中国科学引文数据库(CSCD)和中国社会科学引文索引(CSSCI)两个数据库对我国引用E-print资源的数量、时间、频次, 主要来源站点及其学科情况进行统计分析, 以考察我国对E-print资源的利用情况以及E-print资源在我国学术交流中所起的作用。文章还对国内外主要E-print站点进行了比较, 并结合E-print的交流特征和模式, 提出了促进我国E-print资源建设的措施。

关键词: 学术交流; 引文分析; 电子预印本; e印本文库

中图分类号: G203

文献标识码: A

DOI: 10.3372/j.issn.1674-1544.2010.01.013

Empirical Analysis on Citations to E - prints

Cao Lanfang, Zha Xianjin

(Center for Studies of Information Resources, Wuhan University, Wuhan 430072)

Abstract: This paper collects citation data about E - print from CSCD and CSSCI, and then analyzes the quality, time, degree and the main sources of E - prints cited by scholar, aimed to explore the popularity, importance and effect of E - prints in scholarly communication. Meanwhile, this paper compares the popularity and features of national E - print sites and well - known international sites, and then puts forward some measures to promote the E - print architecture in China, combined with communication characteristics and mode of it.

Keywords: scholarly communication, citation analysis, E - print, E - print archive

1 引言

面对瞬息万变的知识和信息, 如何使科学工作者能在一定范围内合理、高效地共享学术信息资源, 已成为学术界关注的焦点问题。以科技领域为例, 传统的印刷型科技期刊历史悠久, 在印刷型科技期刊上发表科技论文一直以来是科技工作者普遍选择的科技论文共享方式。科技会议录、科技

图书、科技报纸以及由某些核心科学家牵头组织的“看不见的学院(invisible college)”也是科技成果共享的重要方式。但传统环境下科技论文发表与共享长期存在发表周期过长、发行范围受限、运作费用较高等问题。20世纪90年代初, 采用电子预印本(E - print)的方式与学术界同行进行交流逐渐成为一种潮流。E - print具有开放性强、复制成本低、传播速度快、媒体多样化、检索利用方便等特点。它的出现, 加速了学术交流的步伐, 避免

第一作者简介: 曹兰芳(1986 -), 女, 武汉大学硕士研究生, 主要研究方向: 信息资源配置与管理。

基金项目: 教育部人文社会科学重点研究基地重大项目(06JJD870007)。

收稿日期: 2009年9月29日。

了传统的期刊出版发行系统存在的问题。本文将利用中国科学引文数据库(CSCD)和中文社会科学引文索引(CSSCI)两个数据库对我国学术期刊引用E-print资源的情况进行实证分析,考察E-print在我国学术交流中的作用和影响,以及国内主要E-print站点的建设情况与发展策略。

2 数据来源和处理

中国科学引文数据库(CSCD)和中文社会科学引文索引(CSSCI)两个数据库分别收录我国自然科学领域和社会科学领域的引文索引。因此本文利用这两个数据库作为数据源,抽取论文引用E-print资源的数据并进行分析。E-print资源分布在各E-print服务器上,URL没有统一的特征,因此要保证检全率很困难。但大部分E-print资源的URL中含有“eprint”、“preprint”或“e-print”,另外有些作者标引参考文献时会用上述单词表明引用的是E-print文献,所以利用“eprint”和“e-print”进行被引文献检索,能够检索到大部分E-print被引文献。另外,再按照网站的域名特征,逐个检索知名E-print站点的被引资源。这样数据即使不全,也具有一定的代表性。表1列出了中外知名的E-print网站。

表1 中外知名E-print网站

类型	网站名称
学科E-print	arXiv.org, lanl, AKT EPrint Archive, Clinical Medicine NetPrints, Cogprints, CPS, Cryptology ePrint Archive, DoIS, E-BioSci, E-LIS, E-print Network, PhilSic, Public Library of Science, RePEc, SSRN eLibrary, Chemical Physics Preprint Database, CiteBase, 奇迹文库, 中国科技论文在线, 中国预印本服务系统构
机构E-print	ANU E-Prints, CERN Document Server, eprints@iisc, eScholarship Repository, Go8, HofPrints, Open University ePrints, SHERPA
期刊E-print	BBS Online
专类E-print	NDLTD

分别在两个数据库中检索被引来源含有“eprint”、“e-print”、“preprint”或各知名E-print网站的域名特征(如arxiv可以作为arxiv.org的域名特征),并下载数据。数据下载后,还需要进行必

要的处理。首先要去除重复记录和非E-print资源。另外由于CSCD只提供txt和html格式的数据下载,而CSSCI只提供txt格式的下载,所以还需要将数据转换成表格格式,以便分析。关于引文的发表时间和被引时间等数据项也需要自行提取。

3 数据分析

3.1 被引数量分析

CSCD中检索到被引E-print文献4204篇,总被引4321次;CSSCI中检索到被引用E-print文献1202篇,总被引1375次。表2列出了2005年到2008年我国引用E-print文献的篇数和次数。

数据表明,我国自然科学领域从1995年开始引用E-print资源,而社会科学领域从2000年开始,晚了5年,并且无论是总数量还是各年度数量,社会科学领域的引用数量都低于自然科学领域。笔者在SCI与SSCI两个库中利用主要E-print站点的域名特征进行检索,发现存在相同现象。这主要与E-print的发展历史和自然科学、社会科学的学科特点有关。E-print起源于自然科学领域,并且得到较快的发展,但在社会科学领域的发展缓慢,因此自然科学领域的E-print资源就多于社会科学领域,相应的被引用量也较多。另外,由于学科特点的差异,自然学科的学术引用也多于社会学科。根据对各学科学术交流行为的统计数据显示,自然科学领域的期刊年被引次数变化值高于社会科学领域,但累计增长率却低于社会科学领域,这就说明自然科学领域期刊文献的被引数量高于社会科学领域^[1-2]。类推可知,自然科学领域的E-print资源被引数量也应高于社会科学领域。

无论是自然科学领域还是社会科学领域,E-print资源被引量都呈递增趋势,说明其被利用率和关注度越来越高,学术价值越来越大。E-print作为可免费使用的非正式网络学术交流平台,凭借其能与期刊、图书等进行正式交流,与学术会议、学位论文等进行非正式交流,与学术论坛、博客等网络进行学术交流,具有相互联系、有机统一的优点^[3],正得到越来越多学者的青睐,有利于促进我国学术发展。但需要指出的是,虽然E-print

资源被引量逐年上升,但绝对数量仍然很小。CSCD 目前已累积含有引文近 1700 万条^[4],而关于 E-print 的仅 4204 条。说明我国学术界对 E-print 的利用率还是非常低的,这应该与 E-print 文献没有质量保证和在我国宣传力度不够有关。

表 2 我国引用 E-print 文献的篇数和次数

年份(年)	CSCD (篇数/次数)	CSSCI (篇数/次数)	总计 (篇数/次数)
1995	1/1	0/0	1/1
1996	2/2	0/0	2/2
1997	12/12	0/0	12/12
1998	14/14	0/0	14/14
1999	11/11	0/0	11/11
2000	33/36	9/9	42/45
2001	42/42	5/5	47/47
2002	476/482	19/21	495/503
2003	497/498	42/42	539/540
2004	478/481	61/63	539/540
2005	480/488	185/194	665/682
2006	558/567	255/271	813/838
2007	572/592	332/342	904/934
2008	1054/1078	421/428	1475/1506

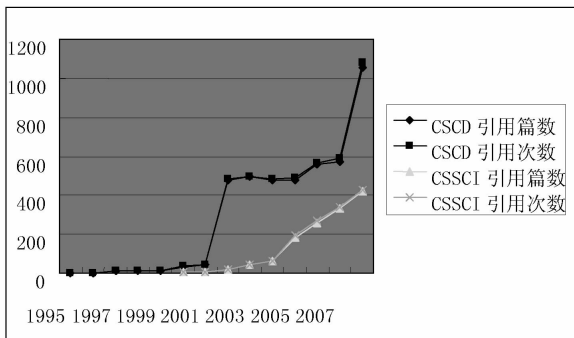


图 1 我国引用 E-print 文献的篇数和次数

3.2 被引时间分析

E-print 资源是由作者以自存档方式提交的电子文献,只需进行简单审稿,用时非常短,再加上网络传播的快速性,E-print 文献从产生到被引用的时间差较短。表 3 列出了 E-print 资源在我国被引用的时差情况。由数据可知,E-print 资源在一年内即被引用的数量最多,这要远远快于期刊文献。有的期刊从审稿到正式发表就需要近一年的时间,甚至更多。另外,从对比表中数据可知,社会科学领域的引用速度快于自然科学领域。这

应与两种科学的特征有关。自然科学领域的实验具有可重复性,且实验周期较长,因此对文献、数据的检验周期较长。而社会科学领域对实验的要求不及自然科学高,因此文献被引用的速度较快。

表 3 E-print 资源在我国被引时差

	当年被引数量	当年被引比率	引用时差众数	引用时差平均数
CSCD	266 篇	6.73%	1 年	4.19 年
CSSCI	195 篇	18.55%	1 年	2.39 年

3.3 被引频次分析

E-print 资源正被越来越多地引用,但绝大部分都仅被引用一次。数据表明,E-print 在自然科学领域被引频次为 1 的文献占总量的 97.86%,而在社会科学领域中占 95.54%。最高被引频次分别为 4 和 5。较低的被引频次说明了 E-print 资源的质量不高,而且未能引起我国学术界的足够重视,在学术交流中发挥的作用有限。E-print 可以弥补传统出版商主导的学术交流方式的不足,但学术评审方面的缺失或不足使其质量得不到保证,因此很少出现高被引文献。另外,网络学术交流虽然日渐受到欢迎,但传统的学术交流模式及交流体系是根深蒂固的,要想立即改变还相当困难。

3.4 主要站点分析

3.4.1 国外主要站点

CSCD 中检索到的 4204 篇文献中有 1440 篇来自 arxiv.org,占总数的 34.25%,大于其他所有站点数据量之和。其次数量较多的站点是 CERN Document Server(247 篇)、Cryptography ePrint Archive(162 篇)、SSRN eLibrary(156 篇)和 lanl(131 篇)。CSSCI 中检索到的 1202 篇文献中有 487 篇来自 SSRN eLibrary,占总数的 40.52%。其次数量较多的站点是中国预印本服务系统(177 篇),RePEc(91 篇),E-LIS(55 篇)和中国科技论文在线(51 篇)。可见我国引用的 E-print 资源主要还是来自国外,特别是 arxiv.org 和 SSRN eLibrary 两个站点。

arxiv.org 是世界第一个,也是目前规模和影响力最大的 E-print 站点。主要收录物理学数学、非线性科学、计算机科学和定量生物学等 5 个学科的 E-print 资源。该系统对学术质量严格控制,

于 2004 年 1 月开始逐步建立审核机制,要求不活跃的研究者在提交 E-print 资源时要得到该领域活跃研究者的认可,这与学术期刊的同行评审制度类似,一定程度上保证了其所收录的 E-print 资源的学术质量。而且文献量大、更新速度快,检索功能强大,因此深受自然科学相关领域学者的喜爱,被引用数量自然也较多。

SSRN eLibrary 是由社会科学电子出版有限公司 (SSEP) 创建的 E-print 站点,主要收录财政金融、会计、法律、经济学和管理学方面的摘要、论文、个人主页和工作论文。它与期刊和研究机构合作,让它们为其提供论文预印本和工作论文的摘要,并设置摘要数据库供用户查询。我国社会科学学者引用这个站点的 E-print 资源最多,说明我国在财政金融、会计、法律、经济学和管理学方面的研究比较多,学术交流氛围较好。

从引文主要来源站点的学科类型可知,我国对 E-print 资源关注较多的学科领域为:物理学、数学、生命科学、计算机科学、图书情报学、经济学、管理学等。而其他很多学科领域则较少或根本没有重视 E-print。

3.4.2 国内主要站点

国内三个主要 E-print 站点是奇迹文库、中国科技论文在线和中国预印本服务系统。奇迹文库是国内最早的中文预印本系统,也是目前国内知名 E-print 站点中唯一由个人创建和维护的。奇迹文库对资源只进行简单的审核,没有任何形式的审稿机制,只要内容完整、不违反基本科学原理的资源都可以存入其中。中国科技论文在线是教育部科技发展中心主办的科技论文网站。该系统注重论文的学术质量,成立了由 35 名中国科学院院士和中国工程院院士组成的顾问委员会提供学术指导,主要负责审查、编辑和发表上传最新的论文,采取“先公开、后评审”的方式。中国预印本服务系统由中国科学技术信息研究所与国家科技图书文献中心联合建设,是国内首家集国内 E-print 文献上下载和国外 E-print 文献下载两项功能于一体的 E-print 网站,不对文章进行学术审核^[5]。

表 4 为国内三大 E-print 网站被引情况。数据显示,我国自然科学领域引用国内 E-print 资

源很少,而社会科学领域的情况则相对好得多,引用国内资源 258 篇,占总量的 1/5。总的说来,我国学术界利用 E-print 资源时,国外资源占绝大多数,自然科学领域几乎全是国外资源。这说明我国 E-print 资源建设水平较低,尤其是在自然科学领域,还需要大力加强。

表 4 国内三大 E-print 网站被引情况

数据库	奇迹文库 (篇)	中国科技论文 在线(篇)	中国预印本 服务系统(篇)	总计(篇)
CSCD	0	1	5	6
CSSCI	30	51	177	258

再对比这 3 个站点:中国预印本服务系统被引量最多,奇迹文库最少,说明中国预印本服务系统建设相对最好,而奇迹文库最差。中国预印本服务系统不进行学术审核,由作者自由提交文稿,而且能进行一站式检索并提供个性化服务,允许注册用户进行个性定制;中国科技论文在线由政府机构创建,采取作者自愿提交和行政干预相结合的方式收集资源,并且实行学术评审,在一定程度上保证了数量和质量,但不足之处是网站界面设计不够友好,内容多且比较零乱,不利于用户使用;奇迹文库由个人创建和维护,经费支持不足,因此规模较小,数量上得不到保证,因而发展相对较弱。

对比国内 3 个站点,再结合国外优秀站点(如 arxiv.org 和 SSRN eLibrary)的成功经验可知,加强我国 E-print 资源建设主要可以从保证数量、提高质量、扩大宣传和优化界面四个方面入手。吸引作者投稿,有了数量保证才能够吸引读者;同时还要采取控制措施,提高资源质量,否则同样难以吸引读者。但数量与质量难以兼顾,所以实际建设中需要掌握一个度,将二者的平衡维持在一定水平之上。除了数量和质量的保证,我国还需要着力扩大宣传,E-print 作为一种重要的非正式学术交流方式,对促进学术进步发挥着很大的作用。但我国目前很多学者对这类资源并不重视,很多人甚至根本不知道什么是 E-print,因此宣传也是发展 E-print 资源的重要策略。最后需要优化界面,分类得当、一站式检索、界面清新有助于读者快速、高效地利用 E-print 资源。

关于提高 E-print 资源数量和保证质量,可

以从 E-print 资源的交流模式与特征方面考虑。E-print 资源由作者自愿上传,不受学术评审的限制,读者可免费获得相关资源并对其内容进行评价,还可以与作者进行交流。省掉学术评审环节是提高学术交流速度的重要措施,但也是导致 E-print 质量得不到保证的根源。对于这种情况,可以参照 arxiv.org 的做法,要求不活跃的研究者在提交 E-print 资源时需得到该领域活跃研究者的认可,这既是一种可保证质量的评审制度,又是一种可提高数量的激励措施,激励研究者提供更多的资源,以成为活跃研究者,从而提高自己在本站的地位,并参与学术评审。E-print 作为一种非正式学术交流方式,应具备扩大交流范围的功能,同时为进一步保证质量和提高数量,可在 arxiv.org 采用的方式上进一步改进,即所有用户均可对已有资源进行学术评价(包括打分和评语),但权重不同,活跃研究者具有最高的权重,提交过资源但不够活跃的研究者具有次高权重,而未提交过资源的普通用户的权重最低。学术评审后质量最高的资源放在显眼的地方,以便读者获取和利用。网站还可以提供重要性排序功能,重要性可由评价的分值乘以权重得出。

4 结 语

E-print 这种非正式交流方式正被越来越多的人接受并利用,尤其是自然科学领域的学者。E-print 不受期刊评审的限制,能够加快学术交流的速度,促进我国学术发展。但也因为同样的原因,其质量得不到保证,所以少有高被引文献。因此,在 E-print 资源建设中需要采取一定措施来保证质量,可以采用先公开后大众评审的方式,并对各领域评分最高的资源进行标识。网络传播中的马太效应可促进大众评审后的高质量资源的广泛利用。arxiv.org、SSRN eLibrary 和 CERN Document Server 是我国学者利用较多的 E-print 资源,而国内的资源则较少被利用。也就是说,我国的 E-print 资源建设还处于较低水平,因此建设高质量的 E-print 资源是当前促进我国学术进步的一

个重要途径。借鉴国外优秀 E-print 成功的经验,并结合我国国情,设计出一套标准,在保证数量的基础上控制好质量,建设出国内高质量的 E-print 资源。因为 E-print 是作者自存档的学术交流方式,所以越多人参与则效用越大,对学术进步的贡献也就越大。而我国很多人到目前为止还并不了解 E-print 的概念和重要意义,加大宣传很有必要。

参考文献

- [1] Qiu Junping, Luo Li. Empirical Analysis of the Regularity in Subject Distribution and Change in Scholarly Communication Behavior Based on JCR (Social Science Edition) [J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(5): 748-754. (in Chinese)
[邱均平, 罗力. 基于 JCR 社会科学版的科研人员学术交流行为的学科分布与变化规律的实证研究 [J]. 情报学报, 2008, 27(5): 748-754.]
- [2] Qiu Junping, Luo Li. Empirical Analysis of the Law of Subject Distribution and Transformation of Scholar Communication Behavior Based on JCR (Science Edition) [J]. Information Science, 2008, 26(2): 174-178. (in Chinese)
[邱均平, 罗力. 基于 JCR 自然科学版的科研人员学术交流行为的学科分布与变化规律的实证研究 [J]. 情报科学, 2008, 26(2): 174-178.]
- [3] 陈光祚. 《e 印本文库(e-print archive)建设与应用》序 [M]// 乔冬梅. e 印本文库(e-print archive)建设与应用——开放存取运动典型策略研究. 北京: 北京图书馆出版社, 2006.
- [4] 中国科学院国家科学图书馆. 中国科学引文数据库介绍 [EB/OL]. [2009-09-19]. http://sciencechina.net.cn/index_more1.jsp.
- [5] Fang Xiangming, Dong Wenyuan. Discussion on Cataloging Policies of Responsible Person in Chinese Machine-readable Catalog Format [J]. Journal of Library and Information Sciences in Agriculture, 2008, 20(1): 96-98. (in Chinese)
[方向明, 董文鸳. 国内外著名学科预印本系统评析 [J]. 农业图书情报学刊, 2008, 20(1): 96-98.]