Characteristics of Open Access Web Citation Network: A Multidisciplinary Study

Kayvan Kousha[1]

27 May 2008

Abstract

More knowledge about Open Access (OA) scholarly publishing on the web would be helpful for citation data mining and the development of Web-based citation indexes. In the current study, five characteristics of 545 OA citing sources targeting OA research articles in four science and four social science disciplines were manually identified, including file format, hyperlinking, Internet domain, language, and publication year. About 60% of the OA citing sources targeting research papers were in PDF format, 30% were from academic domains ending in edu and ac and 70% of the citations were not hyperlinked. Moreover, 16% of the OA citing sources targeting studied papers in the eight selected disciplines were in non-English languages. Additional analyses revealed significant disciplinary differences across science and the social sciences. Overall, the OA Web citation network was dominated by PDF format files and non-hyperlinked citations. This knowledge of some characteristics shaping the OA citation network gives a better understanding about their potential uses.

1   Introduction

The Web is an important source for Open Access (OA) publishing and dissemination of research results. Many have discussed the potential of OA publishing in the scholarly communication cycle (e.g., Harnad 1990; Harnad 1991; Harnad 1999). Others have investigated the citation impact of OA publications in different subject areas (e.g., Antelman 2004; Harnad & Brody 2004; Lawrence, 2001, Kurtz, 2004, Shin, 2003). Since over 90% of journals "have given their official green light to author self-archiving" (Harnad et al. 2004), and an increasing number of authors, journals and institutions are willing to publish their research results online (Swan & Brown 2004; Swan & Brown 2005), a huge amount of citation information, especially from open access Web documents, has become available on the Web. Some researchers have proposed and tested mechanisms for extracting citation information from Web and classified Web citations to validate the web environment as an important source for Scientometrics analysis (e.g., Vaughan and Shaw 2003; Vaughan & Shaw 2005). However, it is not well understood which characteristics have influenced the web citation network and whether disciplinary differences (in science and social sciences) are an important factor in types of characteristics (see below).

The current study focuses on five characteristics of OA scholarly publication (i.e., journal and conference papers, research reports, dissertations) targeting scientific articles in four science and four social science disciplines. In particular, the current study is intended to identify characteristics of OA citing Web documents including file format, hyperlinking, Internet domain, language and publication year and to examine what these characteristics imply for Web citation extraction methods. The results may shed light on the design and development of scientific web mining tools (e.g., Web-based

[1] Department of Library and Information Science, University of Tehran, Iran, Email: kkoosha@ut.ac.ir

citation indexes) and further development of methods capable of capturing and analyzing Web-extracted citation data, especially for OA impact assessment.

1.1 Characteristics of Scholarly Publication on Web

For a long time, bibliometric studies have investigated characteristics of scientific literature based upon the citing and citing relationships within formal literature (i.e., journal article, books). This triggered many studies concerning characteristics of formal citation networks (for a review, see, Moed 2005). The transition of scholarly publishing from print to the Web environment and the increase of open access publishing culture among scholars have developed an evolving Web-based citation network. Several studies have classified Web-based citations to journal articles (Vaughan & Shaw 2003; Vaughan & Shaw 2005; Kousha & Thelwall 2007b); however, the next natural step is to explore the characteristics of this growing citation environment.

Jepsen et al. (2004), for instance, classified the content of 600 URLs retrieved by searching three domain-specific topics related to plant biology in commercial search engines. They found a correlation between PDF files and content classification, indicating that PDF files contain a higher proportion of scientific materials compared with other analyzed Web publication formats. However, the results were only limited to three search topics: "evolution," "genetically modified organisms," and "endangered species" and did not report the influence of other characteristics on scientific publishing on the Web.

Wouters and de Vries (2004) also studied hyperlinking in references, in 38 scientific journals in sociology, library and information science, biochemistry and biotechnology, neuroscience, and the mathematics of computing. They found that the availability of scientific information is not merely determined by the accessibility of Web documents but also by the hyperlinking policy of publishers.

Kousha and Thelwall (2006) studied the characteristics of sources of 1239 formal URL citations targeting 282 research articles pub-

lished in open access library and information science journals. They found that 82% of sources of URL citations were in English; 88% were from full text; 59% were from non-HTML documents and 40% were hyperlinked. However, they only covered the library and information science discipline and one type of citation available on the Web (URL citation). Since there are differences in the extent to which disciplines publish on the web (Kling & McKim 1999; Fry & Talja 2004), more information is needed about the impact of disciplinary differences on characteristics of citation network and scientific publishing on the Web.

The current study covers more disciplines and OA journals and extracts more citations based upon the Web/URL citation method applied in the previous study. Hence, more data is available to address research questions and possible disciplinary differences in science and social sciences.

2   Research questions

The following questions are addressed to investigate common characteristics of citing sources targeting open access research journal articles in science and social sciences.

1. *What are the characteristics of the sources of the Web citations in terms of file format, hyperlinking, Internet domain, publication year and language?*

2. *Are disciplinary differences an important factor in the above characteristics?*

3   Methods

3.1 OA Citing Sources

Since this study is follow-up research, its method is based upon a previous investigation which examined correlation between ISI citations and Google Web/URL citations (Kousha & Thelwall 2007a). In the previous study English language open access peer-reviewed (or editor-reviewed) journals published in 2001were chosen, covering four science disciplines (biology, chemistry, physics, computing) and four social science disciplines (education, psychol-

ogy, sociology and economics). Only research articles were selected, and proportional sampling was applied in each discipline so that journals with more published articles had more sampled articles. This gave 1,650 articles from 108 Open Access (OA) journals. For each article, "Google Web/URL citation" searches (Kousha & Thelwall 2007a) were conducted. These find Web pages that contain either the title of the article or its URL anywhere in the page text (but not necessarily as a link). For this study we only selected sources of Web/URL citations equivalent to formal citation (citations from the reference sections of online documents). Note that citations from cross reference services, Web-based citation indexes and other non-open access citing sources were excluded in this study in order to identify the common characteristics of citation network from open access web documents. Ultimately, 285 and 260 open access citing sources which formally cited the selected articles in the four science and four social science disciplines were respectively selected to examine five pre-defined characteristics.

3.2 Characteristics

As shown in Table 1, five characteristics of each of the 545 open access sources of the formal Web/URL citations (i.e., journal and conference papers, research reports, dissertations) were manually classified and recorded. File formats of the OA citing sources were classified into four sub-classes including PDF, HTML, DOC and Postscript. This shows which file format(s) dominate scientific publication and citation networks on the web. For domain analysis of the citing sources, the proportion of OA citing sources from academic Web spaces with domain names ending in edu or ac (e.g., ac.uk, ac.jp, ac.in) and other non-academic domains (com, org and other) was classified. Although there are many universities and academic institutions that do not use the above academic domains (e.g., Canadian and most European universities), it shows the relative role of universities and academic web spaces in the scientific publication and web citation network and also made the project manageable.

Hyperlinking was another important characteristic which was manually recorded through checking how the citation appeared in the reference section of the citing sources targeting OA articles. This also aids understanding of how the OA Web-based citation network is dominated by hyperlinked or non-hyperlinked citations and to what extend link search methods (i.e., searching the URL of an OA paper) are useful and comprehensive for Web citation data mining.

The languages of the citing sources were also classified into English and non-English. Since many have discussed the over-representation of English language journals in ISI citation databases, the current question might shed light on the role of language diversity in the web scholarly publication and citation network as a useful or trivial pattern. Another aim of the current research is to determine how long it took for an OA article to be formally cited by another Web document and whether disciplinary differences are an important factor for receiving citations on the Web. Thus, the publication year of citing sources during 2001-2005 were identified in different ways, such as referring to the header, footer or footnotes of documents.

3.3 Limitations

Although all the above characteristics were manually identified and recorded through snapshot content analysis of the full text citing documents, there were several limitations. For instance, many OA citing Web documents were automatically created pages from databases (e.g., based upon php and asp programming). The file format of this kind of Web source was classified as HTML. Another limitation relates to the classification of academic domains. There are many universities and academic institutions that do not use the edu and ac academic domains; thus the classification process did not include all academic web spaces. However, there was no practical method to identify and record all the academic domains in the results manually. Regarding languages of the citing sources, two sub-classes were used (English and other languages), thus for practical reasons the study does not reveal the proportion of different non-English languages (i.e., French, Chinese) in the OA citation network. Finally, in some cases it was not possible to discover the exact publica-

tion year of OA citing sources from any of the different methods used, such as referring to the header, footer or footnotes of documents, checking the main (root) URL address of documents and etc. In such cases, the publication year of the citing documents was classified as "not clear".

Table 1. Classification scheme used for exploring characteristics of OA citation networks and related limitations

characteristics	Sub-classes	Limitations
File format	PDF; HTML; DOC; PostScript	Automatically created pages classified as HTML
Domains	Academic domains (edu and ac); Non academic domains (com; other)	Universities that do not use edu and ac not classified
Hyperlinking	hyperlinked citation; non-hyperlinked citation	
Language	English ; other languages	Non-English languages not classified
Pub. year	2001-2002; 2003; 2004; 2005; not clear	publication year for some OA citing sources not identified

4   Results

4.1   File Format

The classification of the 545 OA citing sources targeting OA articles showed that 60% of the OA citing sources were in PDF, 32% were in HTML, 6% in DOC and 2% in the Postscript (PS) file format in the eight studied disciplines. Thus, the findings suggest that OA scholarly publishing on the web is dominated by PDF files. In fact, it seems that the majority of journals, conferences and authors prefer to deposit copies of their articles in PDF format online. Another reason might be related to publishers' policies which is discussed below.

   Figure 1 shows that there are some disciplinary differences in the extent to which web documents publish in various file formats. Most

notably, in biology about 58% of the OA citing sources were from HTML documents. The exploratory study revealed that in biology many citing sources were from HTML papers deposited in PubMed Central[2] which is a digital archive of life sciences in the National Library of Medicine (NLM). In other words, the reason that more HTML citing sources were found in biology could be that more HTML citing sources are available from the above digital archive. Figure 1 also shows that in sociology and economics about 75% and 73% of the OA citing sources were from PDF documents respectively. One explanation for this is the availability of many PDF citing sources from OA journals in sociology and economical reports in economics. For instance, in economics the major source of citation was reports from the World Bank, the World Trade Organisation, and the International Monetary Found, most of them in PDF format. Similarly, in sociology most sources of citation were from PDF journal articles. The results suggest that the policy of key publishers and OA repositories to put OA scholarly documents on the Web can influence the overall results for each discipline. Another conclusion is that for capturing Web citation data, the applied mining tool and methods should be capable of capturing PDF files.
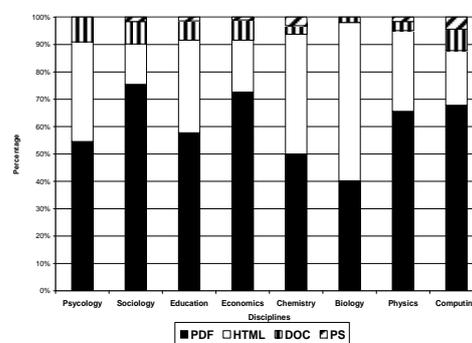

Figure 1. Sources of citations based upon the type of file format

4.2   Hyperlinking

Figure 2 shows the hyperlinking characteristics of the OA citation network. Since the applied

---

[2] www.pubmedcentral.org

method matches a) hyperlinks to the article if the title or URL address of the article appears in the link anchor, and b) the title or URL of the article in the text of other Web pages, even if not hyperlinked (see, Kousha & Thelwall 2007a), it is possible to examine the proportion of hyperlinked and non-hyperlinked citations from the OA citing sources. Only 29% of the citations targeting studied articles in the eight science and social science disciplines were hyperlinked. In contrast, 71% of citations were not hyperlinked (text citation) suggesting that OA scholarly citation network in the studied disciplines was dominated by non-hyperlinked citations.
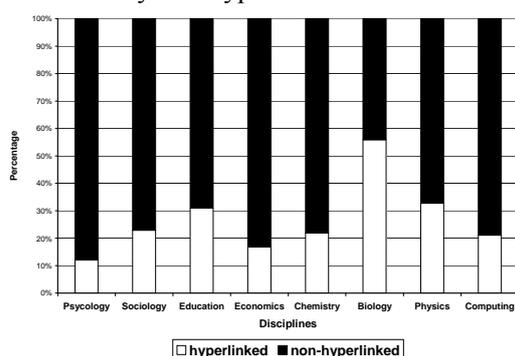

Figure 2. Hyperlinking characteristic of OA citing sources

An important corollary from this study is that using link command searches may not be comprehensive enough for studying research communication on the Web. Although all cited sources were open access and it was supposed that authors might embed URLs for cited items in the references, surprisingly it was found that text citations embedded in PDF citing documents have influenced scholarly publishing on the Web. Most importantly, in biology the percentage for hyperlinked citations (56%) is relatively higher than for non-hyperlinked citations (44%), since many of the citing sources were from HTML papers (58%) deposited to PubMed Central. Consequently, it can be suggested that publishers' policy is an important factor for hyperlinking cited references. For instance, some publishers do not include URL citations in the PDF articles. Another reason may be that when authors convert their papers from MS Word to PDF file format, the hyperlinked citations change to text citations automatically. Although

the latest version of Adobe Acrobat Professional (Version 8) preserves URLs when converting MS Word to PDF file format, some previous versions do not do this.

4.3 Domains

Figure 3 compares the proportion of 545 OA citing sources with domain names ending in edu, ac (e.g., ac.uk, ac.jp, ac.in), com, org and other domains in the eight studied disciplines. As mentioned earlier, many universities do not apply the above academic domains. Thus, the result here only presents an overall view of the significance of the academic web spaces with edu and ac domains. As shown below, except for biology about one third (30%) of the OA citing sources targeting studied papers were from both edu and ac academic domains, which is relatively higher than the org (24%) and com (16%) domains. The result suggests that academic domains have a relatively significant role in formal scholarly communication, although many other academic citing sources were not covered in the current study (e.g., from Canadian universities).
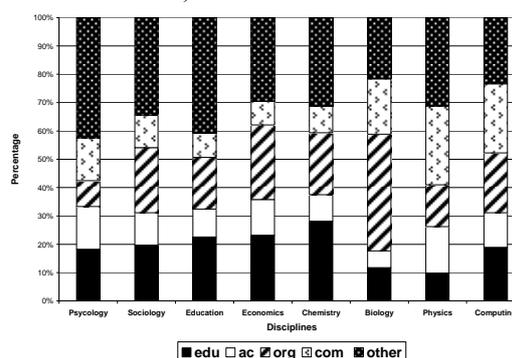

Figure 3. Domains of the OA citing sources

In chemistry many OA citing sources (about 28%) were from American chemical departments/schools which indicates their significant OA publishing contribution. Most notably, in biology findings indicated fewer OA citing sources from academic domains (edu and ac) (18%) and in contrast more citations from the org domain (41%), arising from the dominating role of PubMed Central whose domain name ends with org. Moreover, the proportion of citing sources from the org domain (24%) is considerably higher than that of the com domain

(16%), indicating the significant role of the org domain in online scholarly publishing. One explanation might be that the Internet domain of the key OA repositories and publishers can influence the overall citations from Internet domains.

4.4 Rapid Citations and Language

Another key question is how long it took for an article to be cited by other Web documents. The term "rapid citation impact" is used here to examine the proportion of OA articles which receive a formal citation from other web documents after about two years, 2001-2002, of publishing. In this study journal articles published in 2001 were selected and searches for locating citing sources were conducted in 2005. Thus, it is possible to assess a rapid citing culture in the studied disciplines based upon Web extracted citations. Figure 4 shows that 28% of the OA citing sources targeting studied papers were published during 2001-2002 in the eight selected disciplines. However, in physics about half (46%) of the citing sources were published in 2001-2002, indicating the rapid citation behavior in this discipline, presumably due to preprint sharing. The results support the findings of previous studies that submitting papers to the Arxiv preprint archive lessens the average time between writing a paper and receiving a citation (Brody, Carr & Harnad 2002). The results indicate that publication year characteristic can be extracted from the Web and be used to assess how the citation culture of each discipline differs from that of the others.
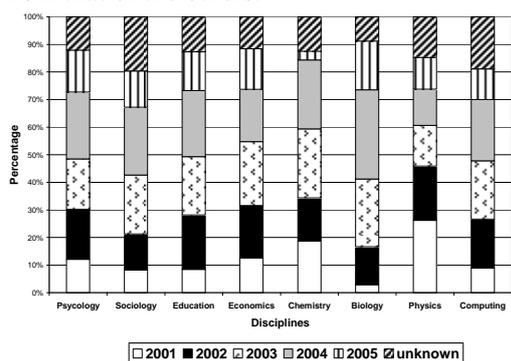

Figure 4. Publication years of the OA citing sources

4.5 Language

The language of the OA citing sources was examined to assess the amount of non-English Web documents. About 16% of the citations targeting the studied papers in the eight selected disciplines were from non-English web documents. Most notably, in computing 22% of the citing sources were in non-English languages. One explanation is that more citing sources were from conference papers and workshops in different languages. Many have discussed ISI limitations in terms of an overrepresentation of the English language; the result suggests that more language diversity indicators can be found online. However, further research is needed to investigate the patterns of language diversity visible through web extracted citations.

5 Conclusions and discussion

5.1 Comparison between Sciences and Social Science

Table 2 compares five key characteristics across the science and social science disciplines in order to examine the impact of disciplinary differences on the characteristics of the OA scholarly publishing in two broad subject areas (science and social sciences). As shown below, more OA citing sources targeting selected articles were in PDF format in the four social science disciplines (65%) than in the science disciplines (56%). Moreover, a higher number of non-hyperlinked citations (79%) was identified in social science than in science (67%). However, it seems that in biology the proportion of HTML citing sources and hyperlinked citation from articles deposited to PubMed (see results) has influenced the overall results in the science disciplines. In other words, if the biology data is ignored, then 61% and 75% of the OA citing sources in the science disciplines were in PDF format and non-hyperlinked respectively. Thus, there are not huge differences between the above characteristics in science and social science.

Domain analysis of the OA citing sources also showed that in the four social science disciplines more sources of citation (33%) targeting studied articles were from the academic do-

mains ending in edu and ac than in the four science disciplines (28%), although the difference is not huge (about 5%). Surprisingly, in science the proportion of non-English citing sources was higher than in the social sciences, suggesting than more languages have influenced web scholarly publishing in the studied disciplines.

Table 2. Comparison of OA citing sources between science and social sciences

	PDF	Domain (edu/ac)	Non-Hyper-linked	Non-English	Rapid citation
Soc. Sci	65.1	33.2	79.3	11.7	27.8
Sci	55.9	28.1	67.1	18.1	30.9

5.2 Predominant characteristics of OA scholarly publishing

The findings suggest that OA scholarly publishing is dominated by PDF format documents and non-hyperlinked citations in the eight studied disciplines. Hence, it seems that the text citation extraction method (text URL or article title searches) might be more useful and effective for studying formal scholarly communication patterns on the web than link extraction methods (e.g., Yahoo! link: searches for article URLs). This is because the link search method can only retrieve citations if hyperlinks to the title or URL address of the article appear in the link anchor.

Further analysis revealed that about 47% and 58% of the OA citing sources were from PDF web documents with embedded text URL citation (non-hyperlinked citations) targeting OA articles in science and social science respectively. In other words, text URL citations were more commonly used in PDF documents. In contrast, hyperlinked citations were more embedded in HTML documents.

Based upon the results, we have a better understanding of the common characteristics of OA scholarly publishing and citation network. This can in practice be used for design and development of web-based scientific data mining tools like autonomous or automatic web-based citation indexes. Finally, exploring the characteristics of formal scholarly communication on the web is becoming more important because increasing numbers of authors, journals and institutions publish and self-archive their research results online. Therefore, similar studies may shed light on how citations that are only found online may be used to help measure online impact of the OA scientific works (i.e., journal articles, conference papers, dissertations, research reports) whose impact was previously impossible to assess through traditional citation indexes like ISI databases.

Acknowledgement

6 References

Antelman, K. (2004), Do Open-Access articles have a greater research impact? College & Research Libraries, 65 (5): 372-382. Retrieved May 4, 2006, from http://eprints.rclis.org/archive/00002309/01/do_open_access_CRL.pdf

Brody, T., Carr, L. & Harnad, S. (2002). Evidence of hypertext in the scholarly archive. Proceedings of ACM Hypertext 2002, Retrieved June 10, 2006, from http://opcit.eprints.org/ht02-short/archiveht-ht02.pdf

Fry, J., & Talja, S. (2004). The cultural shaping of scholarly communication: Explaining e-journal use within and across academic fields. In: ASIST 2004: Proceedings of the 67th ASIST Annual Meeting: Medford, NJ: Information Today Inc., pp. 20-30

Harnad, S. & Brody, T. (2004), Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. D-Lib Magazine, 10 (6). Retrieved May 2, 2006, from http://www.dlib.org/dlib/june04/harnad/06harnad.html

Harnad, S. (1990). Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry. Psychological Science 1: 342–343, Retrieved November, 12, 2004, from

http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad90.skywriting.html

Harnad, S. (1991). Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge. Public-Access Computer Systems Review, 2 (1): 39 - 53.

Harnad, S. (1999). The Future of Scholarly Skywriting, in i in the Sky: Visions of the information future. Retrieved November, 12, 2004, from http://cogprints.org/1698/00/harnad99.aslib.html/

Jepsen E., Seiden P., Ingwersen P., Björneborn L., Borlund P. (2004). Characteristics of scientific Web publications: preliminary data gathering and analysis. Journal of the American Society for Information Science and Technology, 55(14), 1239-1249.

Kling, R., & McKim, G. (1999). Scholarly communication and the continuum of electronic publishing. Journal of the American Society for Information Science, 50(10), 890-906.

Kousha, K. & Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. Scientometrics, 68(3), 501-517.

Kousha, K. & Thelwall, M. (2007a). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis, Journal of the American Society for Information Science and Technology, 57, 1055-1065.

Kousha, K., & Thelwall, M. (2007b). How is science cited on the Web? A classification of Google unique Web citations. Journal of the American Society for Information Science and Technology, 58(11), 1631−1644.

Kurtz, M.J. (2004). Restrictive access policies cut readership of electronic research journal articles by a factor of two, Harvard-Smithsonian Centre for Astrophysics, Cambridge, MA. Retrieved November 13, 2001, from http://opcit.eprints.org/feb19oa/kurtz.pdf

Lawrence, S. (2001), Free online availability substantially increases a paper's impact. Nature, 411, 521. Retrieved November 13, 2001, from http://www.nature.com/nature/debates/e-access/Articles/lawrence.html

Moed, H., F. (2005). Citation analysis in research evaluation. New York: Springer.
Retrieved November 12, 2004 from http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad91.postgutenberg.html/

Shin, E.-J. (2003), Do Impact Factors change with a change of medium? A comparison of Impact Factors when publication is by paper and through parallel publishing. Journal of Information Science, 29 (6), 527 - 533.

Vaughan, L. & Shaw, D. (2003). Bibliographic and Web citations: What is the difference? Journal of the American Society for Information Science and Technology, 54(4), 1313-1324.

Vaughan, L. & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. Journal of the American Society for Information Science and Technology, 56(10), 1075-1087.

Wouters, P. & de Vries, R. (2004). Formally citing the web. Journal of the American Society for Information Science and Technology, 55(14), 1250-1260.