Google Scholar Citations and Google Web/URL Citations: A Multi-Discipline Exploratory Analysis

KAYVAN KOUSHA[1]

MIKE THELWALL[2]

1 Department of Library and Information Science, Visiting PhD Student, School of Computing and Information Technology, University of Wolverhampton, E-mail: kkoosha@ut.ac.ir

2 School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street
Wolverhampton WV1 1ST, UK. E-mail: m.thelwall@wlv.ac.uk

Abstract

In this paper we introduce a new data gathering method "Web/URL Citation" and use it and Google Scholar as a basis to compare traditional and Web-based citation patterns across multiple disciplines. For this, we built a sample of 1,650 articles from 108 Open Access (OA) journals published in 2001 in four science and four social science disciplines. We recorded the number of citations to the sample articles using several methods based upon the ISI Web of Science, Google Scholar and the Google search engine (Web/URL citations). For each discipline, we found significant correlations between ISI citations and both Google Scholar and Google Web/URL citations; with similar results when using total or average citations, and when comparing within and across (most) journals. We also investigated disciplinary differences. Google Scholar citations were more numerous than ISI citations in our four social science disciplines as well as in computer science, suggesting that Google Scholar is a more comprehensive tool for citation tracking in the social sciences and perhaps also in fast-moving fields where conference papers are highly valued and published online. The results for Web/URL citations suggested that counting a maximum of one hit per site produces a better measure for assessing the impact of OA journals or articles, because replicated web citations are very common within individual sites. The results can be considered as additional evidence that there is some commonality between traditional and Web-extracted citations.

1. Introduction

The progressive transition of scientific literature publishing from print to the Web environment has been a key factor in motivating information professionals to explore scholarly communication patterns on the Web, e.g., [1, 2]. In particular, many have considered whether methods of bibliometrics, such as citation analysis, can be applied to the Web environment, for example, [3, 4, 5, 6]. Whilst many early studies analysed links to journal Web sites or online articles [7, 8, 9, 10]; later research tended to extract text-based citations dataset from the Web, but with similar goals [11, 12, 13].
From the early 1990s, many articles have reported the potential of open access (OA) publishing as an emerging scholarly communication phenomenon [e.g., 14, 15, 16, 17].

The next natural step was been to seek evidence for the impact of OA publishing using existing bibliometric techniques [as described in 4] and researchers showed that online availability of articles was associated with higher citation counts in several subject areas [18, 19, 20, 21, 22]. The increasing number of OA journals indexed in the Institute for Scientific Information (ISI) citation databases*, not only indicates their acceptance as a valid outlet for publishing scientific papers, but also encouraged researchers to use ISI citations as a measure of assessment [23] or to compare OA and non-OA journals impact across many disciplines [24].

The ISI's Web of Science has been for a long time the pre-eminent international, multidisciplinary database for citation tracking. Nevertheless, the significant degree of open access publishing in fields such as computer science and physics has allowed some researches to use Web-based repositories to assess the citation impact of articles; the results have typically been compared with ISI-based results for the selected subject area [25, 26] or, on a smaller scale, for an individual journal title [27]. One significant finding was that in computer science, the citations of conference papers seem to be underrepresented by the ISI [25], although it is not clear whether this is desirable, given that the ISI applies quality control mechanisms to select journals for inclusion in their databases, something that does not apply to the web as a whole. In addition to using established web-based repositories, researchers have also developed novel hyperlink-based methods based upon an analogies with citations - both links and citations are inter-document connections, with high numbers of inlinks [28] and citations [29] both being regarded as positive indicators of value - and using commercial search engines for extracting link data [30]. Commercial search engines have also been used as de-facto indexes of the web in order to extract Web citations, which are counted from references in the traditional academic format in web pages [12,13], and URL citations, which are counts of the number of times the URL of a resource is mentioned in other web pages [11]. These studies have tended to compare their results with ISI citations, as a scholarly source with better-known value and validity. In particular, correlation tests have been used as an indirect approach to assess the extent of the agreement between traditional and Web-based citation patterns.

Correlation tests typically take the form of comparing two sets of numbers, such as Web and ISI citations to a collection of journal articles. The test reveals the extent to which larger values from one source associate with larger values from the other source. A high degree of correlation could indicate that one causes or influences the other (e.g., if ISI citations sometimes appear because scholars found references online), or that the two have a common underlying influence (e.g., if both tend to reflect the value of the cited work). This indirect approach is useful as a kind of shortcut to understanding what web measurements may represent by comparing them with better known statistics. Direct approaches, such as a content analysis and interviewing web authors are also needed for the effective interpretation of Web-based variables, however [31]. In particular, we focus on whether Web citation extraction techniques and tools could be considered as substitutes for the ISI counterpart. Disciplinary differences within and between traditional

* At the time of this study more than 200 titles of OA journals were indexed by ISI Web of Science

and Web-based citation counts are important in this context, and hence we place these at the centre of the analysis.

2. Related studies

There is now a considerable body of quantitative research into scholarly use of the web, much of which was reviewed in a recent Annual Review of Information Science and Technology (ARIST) Webometrics chapter [30]. In particular, link analysis is particularly developed field. Much less research has used Web/URL citations, however, for exploring scholarly communication patterns.

Many link analysis studies have been motivated by citation analysis, for example exploring analogies between citations and Web links [32], using the term "sitation" to refer to a cited Web site [6] and defining the "Web Impact Factor" as a Web counterpart of the ISI's Impact Factor for journals [5]. Whilst some information scientists have emphasized the structural similarity between linking and citing [4], others have highlighted the differences between journal citations and Web links [e.g., 33, 34, 35]. Smith [8] was one of the first researchers to use link analysis techniques to examine the relationship between inlinks and ISI Impact Factors for 22 Australasian refereed e-journals, finding no significant association. Similarly, Harter and Ford [7] compared links to 39 scholarly e-journals with ISI citations and found no significant correlation between link counts and ISI impact factors. Although most Webometrics studies have applied quantitative methods (mainly correlation tests), Kim [36] and Herring [37] applied qualitative methods to explore motivations for creating links to journal articles, finding both overlaps with traditional citer motivations and some new electronic medium-specific reasons. The first study to find a significant correlation between the number of links to a journal web site and the associated journal Impact Factor was that of Vaughan and Hysen [9] for ISI-indexed Library and Information Science (LIS) Journals Web sites. Perhaps this study was successful because it was discipline-specific, even though it was dominated by non-OA journals. It was able to take advantage of the fact that most mainstream journals seemed to have deployed an associated web site, which was probably not true at the tie of the early OA studies. Follow-up research confirmed the correlation and showed that journals with more online content tended to attract more links, as did older journal Web sites in both Law and Library and Information Science [10].

In the above studies, "Web links" were the online variable, which was compared with ISI citation counts or journal impact factors. Subsequently, Vaughan and Shaw [12] proposed and applied a new method for extracting citations patterns from the Web, using "Web citations" as measures of impact assessment of journals. They compared ISI citations to LIS journal articles with citations in the Web, using search engine searches to count the number of times each selected journal article was mentioned in web pages. They found significant correlations, suggesting that online and offline citation impacts could be in some way similar phenomena, and hinting that the Web (via search engines) could be a possible replacement for the ISI. In a follow-up study, they found relationships between ISI and Web citations to journal articles in four different areas of science. They also classified Web citations using a predefined scheme to examine the portion of Web citations that reflect the intellectual impact of the articles [13]. Most selected journals were traditional ISI journals with independent Web sites but were not open access. They suggested that Web and ISI citation counts were measuring the same things in assessing the impact of journals or their papers.

URL citation analysis, counting the number of times the URL of an OA article is mentioned on the web, is different to both link analysis and citation analysis; although an URL citation is also a link when the URL is also a hyperlink. One advantage of URL citation analysis over web citation analysis is that URLs are unique, whereas paper titles are not. A disadvantage is that both links and citations may exclude a visible URL citation and so URL citations capture only a proportion of times an article is referred to online. In fact URL citations, links and web citations all overlap to some extent, but not completely (although the MSN Search can simultaneously capture hyperlinks and URL citations [38]. A comparison of peer-reviewed open access LIS journal articles found a significant correlation between ISI and URL citation counts and also between average numbers of ISI and URL citations [11]. A classification of URL citations in this study estimated that 43% reflected citation-like intellectual impact and a further 18% represented informal scholarly uses.

There are relatively few comprehensive studies across several subject areas comparing conventional citations (e.g. ISI citations) with citations from Web-based citation indexes at the article or journal level. Goodrum [25], for instance, compared citation patterns in online computer science papers indexed in CiteSeer with citations from the ISI. They found that conference papers in computer science were relatively more frequently cited online and much less by ISI articles. Zhao & Logan [26] conducted a similar study in the XML research area and found that CiteSeer, provided more citations than the ISI for this relatively new and fast moving field. Bauer & Backlash [27] compared the citation counts provided by the ISI Web of Science, Elsevier's Scopus abstract and indexing database, and Google Scholar for articles from the Journal of the American Society for Information Science and Technology (JASIST) published in 1985 and 2000. For articles published in 2000, Google Scholar provided statistically significant higher citation counts than either the Web of Science or Scopus, whilst there was no significant difference between the Web of Science and Scopus.

The citation facility of Google Scholar (http://scholar.google.com) is a potential new tool for bibliometrics. Launched in November 2004, Google Scholar claims to include "peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations. Google Scholar helps you identify the most relevant research across the world of scholarly research" [39]. Perhaps some of these documents would never been indexed by search engines such as Google, so they would be "invisible" to web searchers, and some would be similarly invisible to Web of Science users.

A number of authors have noted some problems with the early Google Scholar, especially uneven coverage of scholarly publishers' archives and false drops [40, 41]. Nevertheless it has also been heralded as valuable because of its coverage of academic information from many publishers, such as the ACM, Annual Reviews, arXiv, Blackwell, IEEE, Ingenta, the Institute of Physics, NASA Astrophysics Data System, PubMed, Nature Publishing Group, RePEc (Research Papers in Economics), Springer, and Wiley Interscience [42]. Many Web sites from universities and nonprofit organizations are also included; most notably OCLC Open WorldCat, with millions of bibliographic records

[42]. The question is can researchers and students, especially those who have not access to conventional fee-based citation indexes, such as Web of Science and Scopus, use the Google Scholar for locating scholarly information? Despite all the reviews, no previously study used extensive data collection method and analysis in different science and social science disciplines to answer the following question. For instance, few evidences about ISI and Google Scholar citation impact is available based upon the a single year of a journal articles searches in both databases related to the specific discipline [27] and such results can be barely generalized to at least an individual discipline.

3. Research questions

We address three questions to compare ISI and different types of Web-extracted citation patterns at the individual article and journal level. Note that correlation tests preformed in this study are used as an indirect approach for interpreting Web-extracted citation counts and further study is needed for direct interpretations of Web citations. We will also examine the apparent disciplinary difference within and among subject areas in terms of the relationship between traditional and Web-based citation patterns as well as the number and average of citations.

- Is there a correlation between ISI citation counts and either Google Scholar or Google Web/URL citation counts for the articles in OA science and social science journals?
- As above but using average citation counts per article across individual OA journals.
- Are there strong disciplinary differences between conventional and Web-extracted citation patterns within and between individual science and social science disciplines?

4. Method of Study

4.1. Discipline, Journal and Article Selection

For the purpose of this study, OA journals are freely accessible English language journals available on the Web (in electronic only or both in electronic and print formats) with articles that have undergone some kind of peer or editorial review process, and irrespective of whether the electronic publishing is the primary or secondary medium for the journal [43]. We needed OA journals that had been published at least since 2001 in order to allow a significant time window in which to attract citations. An initial study based upon the Directory of Open Access Journals (www.doaj.org) and an ISI essay on the impact of open access journals [24] showed that there were few science and social science disciplines having enough refereed or editor-reviewed open access journals published in 2001. This low rate of OA journal publication in many areas is a limitation of our study and is mentioned again in the discussion. We used both of the above sources and other related directories to locate OA journals for this study. Ulrich's Periodical

Directory (2004) was consulted for official journal Web site URLs and the availability of OA journals in electronic only or both in electronic and print formats. The study only used the official Web sites of OA journals (the journal publisher's Web site) for recording article URLs. Therefore, URLs of articles in mirror sites were not examined. Journals which didn't have an independent Web site were also included in this study, because the data collection method (Google searches) were preformed for locating Web/URL citations to journal articles in the text of other Web pages, not links to whole journal Web sites. If a journal Web site was in an individual HTML 'frame' and all its articles had the same URL then we excluded it from this study. We also excluded journals ceasing publication before, or with publication beginning after, 2001. Our final sample included 108 open access journals, 55 of which were indexed in the ISI Web of Science at the time of this study (Table 1).

The factors that were considered when selecting the science and social science disciplines included the number of refereed or editor-reviewed OA journals in each discipline, the accessibility of journal web sites for data gathering. We selected biology, chemistry, physics and computer science to represent a range of (hard) sciences and economics, education, sociology and psychology to represent a range of social sciences. The selection of four science and four social science disciplines allowed comparisons between broadly similar disciplines as well as between distinctly different ones.

For each selected journal, we applied a stratified sampling method for a systematic selection of journal articles (omitting reports, editorials, and book reviews). The exact title of each research article was recorded, along with its HTML or PDF URL. For a statistically representative selection, in each discipline we took a random sample proportional to the total number of articles in each journal. Consequently, in each discipline journals with more published articles had more articles in our final sample. Ultimately, 1650 research articles were selected from 108 journals in eight disciplines. Using Ulrich's Periodical Directory and information in journal Web sites, we found that 52% (56 of 108) of the selected journals were available in both print and electronic formats and 48% (52) were exclusively available online (Table 1). Appendix A shows the journals selected for this study.

Table 1. Sample statistics for the sample collection of journal articles.

Selected Disciplines	Number of selected OA journals	Number (%) of articles sampled	Number of journals indexed in the ISI Web of Science	Number (%) of journals available in both print and electronic	Number (%) of journals available exclusively online
Biology	21	325 (19.7%)	17	13 (62%)	8 (38%)
Chemistry	15	325 (19.7%)	15	12 (80%)	3 (20%)
Physics	16	325 (19.7%)	12	11 (69%)	5 (31%)
Computing	12	183 (11.1%)	5	9 (75%)	3 (25%)
Education	17	185 (11.2%)	2	5 (29%)	12 (71%)
Economics	11	134 (8.1%)	1	2 (18%)	9 (82%)
Sociology	7	70 (4.2%)	1	2 (29%)	5 (71%)
Psychology	9	103 (6.2%)	2	2 (22%)	7 (78%)

Total	108	1650 (100%)	55	56 (52%)	52 (48%)

4.2. ISI, Google Scholar and Google Web/URL Citation Counts

Each of the 108 journals was searched for in the ISI Web of Science in order to examine if they had been indexed. For the 55 ISI-indexed journals (at the time of this study), the number of citations to each article in the sample was recorded. Since many open access journals weren't indexed (53 titles) or were indexed after 2001, their names were searched for in the "Cited Reference Search" field of the ISI Web of Science as an alternative way to find out the number of citations their 2001 articles had received from other ISI-indexed journal articles. In order to prevent possible errors due to similar abbreviations for different journals, the first author name and volume of each retrieved article was checked against the original OA article in the sample. This method is similar to that applied by Vaughan & Shaw [13] for ISI indexed journals and latter by Kousha & Thelwall [11] for journals not indexed by ISI. For the purpose of this study, both searches were limited to citations to year 2001 articles and journal names were truncated if necessary.

For Google Scholar citation counts, we searched the titles (taken from journal web site tables of contents) of all 1,650 sampled articles as phrase searches in the main Google Scholar search page. We found that some titles with mathematical or chemical formulas did not retrieve any matches if their complete titles were used. Thus, it was necessary to omit a portion of some article titles (especially in physics, chemistry and biology) during the search process in order to generate effective searches. We manually checked the search results against the original citation information to avoid false matches and to remove any duplicate citing documents. We then recorded the number of Google Scholar citations by clicking the "cited by" option below each retrieved record after omitting the obvious false drop.

Google was chosen for extracting Web/URL citation counts because results of previous studies showed that it has provided the most comprehensive [44] and the most stable search results over time [13, 45]. Google has good coverage of HTML and non-HTML documents and supported the syntax necessary for extracting both Web and URL citations at the time of this study, as described below. Nevertheless, the results of Google do not represent the whole web [46], only the portion of the web that it has crawled and reports for user searches [47, 48].

For what we call Google Web/URL citations, methods from two previous studies were used. The article title phrase search method of Vaughan and Shaw [13] for retrieving Web citations and the URL search method used by Kousha and Thelwall [11] for locating URL citations were combined. The method applied, as shown below for the PDF version of an article from the Journal of Chemical Sciences, matches (1) hyperlinks to the article if the title or URL address of the article appears in the link anchor, and (2) the title or URL of the article in the text of other Web pages, even if not hyperlinked. We used –site: in order to exclude Web/URL citations from the same journal Web site. For very general short article titles we added author(s) or the journal name to our syntax to avoid retrieving unwanted results.

"Enantioselective solvent-free Robinson annulation reactions" OR
www.ias.ac.in/chemsci/pdf-Jun2001/Pc3049.pdf

-site:ias.ac.in/chemsci

For articles available in HTML and PDF format, both URLs were combined in the searches through the Google OR operator. No previous study has used this kind of data collection method, and it is clearly more comprehensive than either of the two methods that it combines. We believe that our method probably includes the vast majority of all types of Google-indexed Web citations or links, but further research would be needed to verify this.

We selected the option “repeat the search with the omitted results included”, if it was displayed, to retrieve total number of results in Google. Note that all the ISI, Google Scholar and Google Web/URL searches in this study were conducted for each discipline during the relatively short period September-October 2005 in order to minimize the potential impact of time on increasing the number of citation counts, and of variations in Google’s web coverage.

4.3. Google unique and total Web/URL citations

We found that selecting the option “Repeat the search with the omitted results included” at the bottom of Google searches sometimes retrieved many results with similar contents from individual sites. In fact, Google total results often contain a separate hit for the main entry, the abstract, the PDF file and (if available) the HTML file of each article. Although URLs of such hits are slightly different, they direct the users to the same article in the site. We believe that these results should be considered redundant. Since this is a relatively new issue, we decided to record both kinds of Web/URL citation counts and to investigate which one has a higher correlation with ISI citation patterns at the individual article and journal levels. The large differences between mean and median Google unique and total Web/URL citations can be seen in Table 2. In summary, for the purpose of this study we defined Google unique Web/URL citation counts as the number of Web/URL citations per one site. Since Google often gives two hits per site, this number was manually calculated based upon including only one result per site. The number of unique Web/URL citations was conveniently calculated by omitting the indented Google results. We also recorded Google total Web/URL citation counts as the total number of Web/URL citations retrieved by Google, after omitting any identified false matches.

5. Findings

The correlation tests in Table 2 were calculated for each discipline using individual sampled papers as the unit of data collection and data analysis. This part of the study presents a broad view of disciplinary differences between counts of traditional and Web-extracted citations. Following Vaughan and Shaw [12], Pearson correlation tests were preformed if the frequency distributions were not very skewed; otherwise Spearman correlation tests were applied.

Table 2 Correlations between ISI, Google Scholar and Google Web/URL citation counts to OA articles and descriptive statistics for each eight studied disciplines.

No. of articles sampled for correlation study	**Total Google Web/URL citation** Mean, Median, Total	**Unique Google Web/URL citation** Mean, Median, Total	**Google Scholar citation** Mean. Median, Total	**ISI citation** Mean, Median, Total	ISI and total Google web/URL citations	ISI and Unique Google web/URL citations	ISI and Google Scholar Citations	Selected disciplines
325	41.046,22.000, 13340	11.492, 8.000, 3753	4.855. 2.000 1578	5.701, 3.000 1853	0.259**	0.325**	**0.825****	Biology
325	9.160, 5.000 2977	3.670, 3.000 1193	1.027, 0.000 334	2.375, 1.000 772	0.285**	0.324**	0.553**	Chemistry
325	20.604, 11.500, 6717	6.901, 5.000, 2250	3.742, 1.000, 1220	3.816, 1.000, 1244	0.415**	0441**	0.672**	Physics
183	99.907, 49.00, 18283	21.486, 15.000, 3932	10.978, 2.000, 2009	5.453, 1.00, 998	**0.702****	**0.746****	0.815**	Computing
185	47.940, 28.000 8869	19.129, 12.000, 3539	2.821, 1.000, 522	0.556, 0.000, 103	0.412**	0.438**	0.551**	Education
134	74.776, 45.500, 10020	22.522, 15.000, 3018	5.119, 2.000, 686	0.666, 0.000, 89	0.467**	0.524**	0.574**	Economics
70	78.557, 27.5, 5499	19.171, 13.000, 1342	3.885, 1.000, 272	1.771, 0.000, 124	0.519**	0.616**	0.766**	Sociology
103	40.135, 13.000 4134	14.330, 8.000 1476	2.524, 1.000 260	1.262, 0.000, 130	0.055	0.054	0.563**	Psychology

** Significant at the p = 0.01 level.

5.1. ISI citations correlate with Google Scholar and Web/URL citation counts

As shown in Table 2 there is a significant correlation between the ISI and Google Scholar citation counts in all studied disciplines ($p < 0.01$). Although the correlations for science disciplines are higher than for social science disciplines, especially for biology and computer science, the exceptions are sociology (high) and chemistry (low). A possible explanation for the differences in correlation coefficients is better coverage of science journals by the ISI and Google Scholar than social science journals; this will be discussed in the next section. Nevertheless, this is reasonable evidence that scholarly OA journal articles with more citations in the ISI database also have more citations reported by Google Scholar for science and the social sciences, although there may be disciplines that we have not studied for which this is not true. Moreover, since both ISI and Google Scholar citation counts measure the same formal patterns of scholarly communication (conventional citations), this is also reasonable evidence that they are essentially assessing something very similar, which might for convenience be called the intellectual impact of the work, although the exact meaning of citation counts is far more complex than this [29].

We can see the same relationship, albeit weaker, between ISI and Google unique citation counts as well as Google total citation counts for seven disciplines (excluding psychology) at the $p = 0.01$ level. In both cases there are higher correlations for computer science and sociology and lower for biology and chemistry. The higher Google Scholar correlations for science disciplines are not reflected in the Web/URL citation correlations; three social science disciplines (sociology, economics and education) have higher Web correlations (Google unique and total Web/URL citations) than three of the hard science disciplines (biology, chemistry and physics). Finally there are higher correlations between ISI citations and Google unique Web/URL citations than Google total Web/URL citations in both science and social science disciplines (excluding psychology), supporting our argument above that unique (one per site) Web/URL citation counts in Google search results are a better scholarly measure than total Web/URL citations.

5.2. Social science articles receive more formal citations on the Web than ISI

Table 2 shows that the mean and total number of ISI citations to sampled articles in three science disciplines (biology, chemistry and physics) is higher than Google Scholar citations. Moreover, the median of ISI citations to sampled articles in biology and chemistry is higher than Google Scholar citations and physics has an equal median for both kinds of citation counts. In contrast, in four social science disciplines (sociology economics, psychology and education) the mean, median and total number of ISI citations is much lower than Google Scholar citations suggesting that Google Scholar may have particularly good coverage of sources for citations in the social sciences, but may be slightly weak for the sciences. In computer science the mean, median and total number of Google Scholar citations is much higher than ISI citations (about double).

One explanation is that in computer science there is an existing established web citation database (CiteSeer/ResearchIndex) and in computer science conferences are very important and their proceedings are frequently made available online, where Google Scholar may find them. In physics, conferences are not the major dissemination mechanism but open access publishing is of preprints is common, via ArXiv.org. Table 2 also shows that the mean, median and total number of Google Web/URL citations (unique and total) is much higher than both ISI citations. This corroborates the work of Vaughan and Shaw [13] for 114 biology, genetics, medicine, and multidisciplinary science journals.

5.3. Correlations between ISI and Web citation counts for each journal

In the previous section we calculated correlation coefficients for each discipline using individual sampled papers as the unit of data collection and data analysis. In this section we break down the data further and report correlation tests for ISI and three different Web-extracted citations counts (Google Scholar, unique and total Web/URL citations) for each journal in each eight disciplines, again using the papers in our sample as the unit of data analysis and data collection. Note that journals with less than 10 articles in our sample were excluded for the reliability of correlation tests between variables (26 journals). Table 3 shows the percentage of significant correlations between the ISI and three types of Web-extracted citations for each discipline. For instance, Table 3 shows that there is a significant correlation between ISI and Google Scholar citation counts for 80% (16 out of 20 titles) of the biology journals. Table 3 also shows that the average percentage of significant correlations between ISI and three Web-extracted citations in all studied disciplines is higher for Google Scholar citations (66.5%), than unique Web/URL citations (30.4%), and total Web/URL citations (19.4%).

Table 3. Percentage of significant correlations between ISI and Web–extracted citation counts for the journals in science and social science disciplines.

Disciplines	Significant	Significant	Significant correlation	Number of

	Correlation between ISI and Google Scholar Citation %	correlation between ISI and Google unique Web/URL Citation %	between ISI and Google total Web/URL Citation %	journals for correlation study
Biology	80	25	20	20
Chemistry	76.9	15.4	15.4	13
Physics	62.5	43.7	43.7	16
Computer	75	50	25	8
Education	41.6	8.3	8.3	12
Economics	66.6	33.3	0	3
Sociology/Psychology	70	40	30	10
Average percentage of significant correlations	66.5	30.4	19.4	82

Table 4 gives a more general view of differences between the percentage of significant correlations for two major science and social science disciplines.

Table 4. Percentage of the significant correlation between the ISI and web–related citations data for the journals in science discipline

	Percentage of Sig. Correlation between ISI and Google Scholar Citation	Percentage of Sig. correlations between ISI and Google unique Web/URL Citation	Percentage of Sig. correlation between ISI and Google total Web/URL Citation	Number of journals
Science	73.6%	33.5%	26.02%	57
Social Science	59.4%	27.2%	12.8%	25

5.4. Average ISI citations correlate with average Google Scholar and Web/URL citations

Correlation tests were also performed using individual journals as the unit of data analysis in each selected discipline; between the average number of ISI citations and the average number of Google Scholar citations and the average number of Web/URL citations for each of the 108 journals (Table 5). For each variable we calculated the total number of citations (ISI, Google Scholar and Web/URL citations) a journal received divided by the number of papers in the sample set. As shown in Table 5, there is a highly significant correlation between the average number of ISI citations and the average number of Google Scholar citations in all the disciplines at the p = 0.01 level. It is interesting that there is a relatively higher correlation for OA journals in science than in the social science disciplines. We merged journals of two subject areas, sociology and psychology, because there weren’t enough journals (data points) in each of them for a correlation test. Thus, it seems that OA journals having higher average ISI citations also have higher average Google Scholar citations. Hence Google Scholar is a promising tool for measuring the intellectual impact of OA journals as an alternative to conventional citation indexes.

We also found significant correlations between the average number of ISI citations and the average number of Google unique Web/URL citations (as we defined above) at the journal level for each discipline (Table 5), but significantly lower than the correlations between average ISI and Google Scholar citations reported above. We found significant correlations between average ISI and average Google total Web/URL citations at the journal level for four disciplines.

Table 5 Correlations between average ISI and average Google Scholar and Google Web Citation Counts to OA journals.

No. of OA journals in 2001	Correlation between average ISI and average Google total web/URL citations	Correlation between average ISI and average Google unique web/URL citations	Correlation between average ISI and average Google Scholar Citations	
21	0.248	0.622**	0.938**	Biology
15	0.443	0.721**	0.744**	Chemistry
16	0.503*	0.547*	0.926**	Physics
12	0.782**	0.831**	0.880**	Computer
17	0.597*	0.611**	0.782**	Education
11	0689*	0.734*	0.655**	Economics
16	0.470	0.540*	0.789**	Sociology/Psychology

* Significant at the $p < 0.05$ level.
** Significant at the $p < 0.01$ level.

5.5. Journal Impact Factors correlate with average ISI, Google Scholar and Web/URL citation counts

We calculated correlations between ISI Journal Impact Factors (JIF) and average Google Scholar/Web citation counts for 47 journals. As mentioned above, of 108 selected journals in this study, 55 titles were indexed in the ISI Web of Science. We found that only 47 titles had impact factors in the 2004 edition of ISI Journal Citation Report (JCR) at the time of this study. The year 2004 Impact Factors are calculated based upon cites in 2004 to articles published in 2003 and 2002. We found significant correlations between JIFs and average Google Scholar citations (r = 0.624**), average Google unique Web/URL citations (r = 0.475**) and average Google total Web/URL citations (r =0.387**) respectively. The results show that journals with higher ISI Impact Factors also had higher average Web-extracted citations.

6. Discussion and conclusions

We found a significant association between the ISI citations and both Google Scholar and Google Web/URL citations to open access scholarly journals in science and social science disciplines, indicating that conventional and Web-based citations patterns are likely to be measuring similar things and have the potential to be used for impact assessment, if further research successfully corroborates these findings and investigates reasons for the differences and reliability issues. We also found a relatively stronger relationship between the number of and average ISI and Google Scholar citations than Google Web/URL citations in nearly all cases at the article and journal level. One explicit and clear explanation for such relationship is that both ISI and Google Scholar are measuring formal scholarly patterns, equivalent to formal citations and that Web/URL citations include types of informal citation in addition to the formal ones. It is not clear, however, which is the better type of measure for research impact. For example, if many Web/URL citations represented genuine uses of research (e.g. in education or industry) then this could be seen as desirable, whereas if most Web/URL citations were in replicated library lists, then this could be seen as problematic.

Separating out the unique and total Web/URL citation counts in our Google search results showed that there are relatively higher correlations between the number of, and average, ISI and Google unique Web/URL citations, suggesting that counting a maximum of one match per site produces improved results. The fact that through Google total results we retrieve many multiple links which direct the users to the same place in the site can be considered as the main reason for the relatively stronger relationship between ISI and Google unique Web/URL citation counts in all our correlation tests at the journal and article level.

Our new Web/URL citation method is, in theory, more comprehensive than either web citation or URL citation. Nevertheless, it has some practical drawbacks. As we have used it, it requires some manual labor to ensure that only one citation per web site is allowed, although this could be automated in Google [49]. Moreover, the method required modification for some journals because of too-long URLs, which creates a potential unfairness. Finally, it is not yet clear that the most comprehensive solution is the best, especially if many of the additional citations may come from undesired sources.

Why is there a relatively stronger correlation between ISI and Google Scholar citation counts in science disciplines than social science at the journal and article level? It may be relevant that about 77% (49 of 64) of selected journals in science disciplines and only 13% (6 of 44) of social science journals were indexed by the ISI at the time of this study. The descriptive statistics for ISI and Google Scholar citation counts (Table 2) shows that in three pure science disciplines (biology, chemistry and physics, but excluding computer science) the distribution of citations is relatively less skewed than in social science. Consequently, it may be that higher coverage of citation information in both ISI and Google Scholar is the important factor for higher commonality between citation patterns in science disciplines. This speculation is supported by the fact that, in contrast to three pure science disciplines, in four social science subject areas the number, the mean and median Google Scholar citation counts is remarkably higher than ISI citations (Table 2). This suggests that Google Scholar is a more comprehensive tool for citation tracking in social science in this study. However, the quality of sources of citations (citing documents) retrieved by Google Scholar is important factor to take into account (as for Web/URL citations), and future research must address this complex issue. In addition, there are disciplinary differences in research which lead to varied emphasis on things like electronic publication, books, journals and conferences [2, 50] and variations in usage patterns for similar electronic resources, including e-journals [51]. It may be that the web contains objects of value in the social sciences, such as course reading lists, that would not be used or valued in the sciences if research is less directly tied to teaching. A corollary from this, and the fact that our disciplines were effectively a convenience sample, self-selected by volume of OA journal use, is that it would be unwise to assume that OA will become the norm throughout academia. Whilst the methods here and particularly Web/URL citation advantage web-published sources, it would be unfair to use them to compare non-OA journals, even though (non-OA) web publishing seems to be standard now for the major academic publishers.

References

1. J. Fry, The cultural shaping of ICTs within academic fields: Corpus-based linguistics as a case study. Literary and Linguistic Computing, 19(3), 303-319, 2004.
2. R. Kling and G. McKim, Scholarly communication and the continuum of electronic publishing. Journal of American Society for Information Science, 50(10), 890-906, 1999.
3. T. C. Almind and P. Ingwersen, Informetric analyses on the World Wide Web: Methodological approaches to “Webometrics”. Journal of Documentation, 53(4), 404-426, 1997.
4. C. Borgman and J. Furner, Scholarly communication and bibliometrics. Annual Review of Information Science and Technology, 36, Medford, NJ: Information Today Inc., pp. 3-72, 2002.
5. P. Ingwersen, The calculation of Web Impact Factors. Journal of Documentation, 54(2), 236-243, 1998.
6. R. Rousseau, Sitations: An exploratory study. Cybermetrics, 1(1), 1997, Retrieved November 14, 2001, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html/
7. S. Harter and C. Ford, Web-based analysis of E-journal impact: Approaches, problems, and issues, Journal of the American Society for Information Science, 51(13), 1159-76, 2000.
8. A.G. Smith, A tale of two Web spaces: Comparing sites using Web impact factors. Journal of Documentation, 55(5), 577-592, 1999.
9. L. Vaughan and K. Hysen, Relationship between links to journal Web sites and Impact Factors. Aslib Proceedings: New Information Perspectives, 54(6), 356-361, 2002.
10. L. Vaughan and M. Thelwall, Scholarly use of the Web: What are the key inducers of links to journal Web sites? Journal of the American Society for Information Science and Technology, 54(1), 29-38, 2003.
11. K. Kousha and M. Thelwall, Motivations for URL citations to open access library and information science articles. Scientometrics, to appear, 2006.
12. L. Vaughan, and D. Shaw, Bibliographic and Web citations: What is the difference? Journal of the American Society for Information Science and Technology, 54(4), 1313-1324, 2003.
13. L. Vaughan, and D. Shaw, 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, 2005.
14. S. Harnad, Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry. Psychological Science 1: 342 – 343, 1990, Retrieved November, 12, 2004, from http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad90.skywriting.html/
15. S. Harnad, Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge. Public-Access Computer Systems Review, 2 (1): 39 – 53, 1991, Retrieved November 12, 2004 from http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad91.postgutenberg.html/
16. S. Harnad, The impact of electronic journals on scholarly communication: A citation analysis. The Public-Access Computer Systems Review, 7, 1996, Retrieved November 13, 2001, from http://info.lib.uh.edu/pr/v7/n5/hart7n5.html/

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

18. K. Antelman, Do Open-Access Articles Have a Greater Research Impact? College & Research Libraries, 65(5): 372-382, 2004.

19. S. Harnad, T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E. Hilf, The access/impact problem and the green and gold roads to open access. Serials Review 30, 2004, Retrieved November, 12, 2004, from http://www.nature.com/nature/focus/accessdebate/21.html/

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

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

22. E.-J. Shin, 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, 2003.

23. T. Brody, H. Stamerjohanns, F. Vallières, S. Harnad, Y. Gingras, and C. Oppenheim, The effect of open access on citation impact, 2004. Retrieved November 13, 2001, from http://www.ecs.soton.ac.uk/~harnad/Temp/OA-TAadvantage.pdf/

24. ISI press release essay on the impact of open access journals: A citation study from Thomson ISI. Retrieved November 13, 2004, from http://www.isinet.com/oaj/

25. A.A. Goodrum, K.W. McCain, S. Lawrence, and C.L. Giles, Scholarly publishing in the Internet age: a citation analysis of computer science literature. Information Processing & Management, 37(5), 661-676, 2001.

26. D. Zhao and E. Logan, Citation analysis using scientific publications on the Web as data source: A case study in the XML research area. Scientometrics, 54(3), 449-472, 2002.

27. K. Bauer and N. Bakkalbasi, An Examination of Citation Counts in a New Scholarly Communication Environment. D-Lib Magazine, 11(9), 2005, Retrieved December 23, 2005, from http://www.dlib.org/dlib/september05/bauer/09bauer.html /

28. S. Brin and L. Page, The anatomy of a large scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7), 107-117, 1998.

29. H., F. Moed, Citation analysis in research evaluation. New York: Springer, 2005.

30. M. Thelwall, L. Vaughan, and L. Björneborn, Webometrics. Annual Review of Information Science and Technology, 39, Medford, NJ: Information Today Inc. 81-135, 2005.

31. M. Thelwall, Interpreting social science link analysis research: A theoretical framework. Journal *of the American Society for Information Science and Technology*. 57(1), 60-68, 2006.

32. A.G. Smith, Web links as analogues of citations. Information Research, 9(4), 2004, Retrieved March 20, 2005, from http://informationr.net/ir/9-4/paper188.html

33. L. Björneborn and P. Ingwersen, Perspectives of Webometrics. Scientometrics, 50(1), 65-82, 2001.

34. L. Egghe, New informetric aspects of the Internet: some reflections - many problems. Journal of Information Science, 26(5), 329-335, 2000.

35. W. Glänzel, On some on some principle differences between citations and sitation links. A methodological and mathematical approach. Nerdi lecture delivered at NIWI, KNAW, Amsterdam, on 13 February, 2003. Updated version of a paper presented at the 6th Nordic Workshop on Bibliometrics, Stockholm, October 4-5, 2001.

36. H.J. Kim, Motivations for hyperlinking in scholarly electronic articles: A qualitative study. Journal of the American Society for Information Science, 51(10), 887-899, 2000.

37. S.D. Herring, Use of electronic resources in scholarly electronic journals: A citation analysis. College and Research Libraries, 63(4), 334-340, 2002.

38. D. Stuart, Personal communication, 2006.

39. About Google Scholar, Retrieved December 12, 2005, from http://scholar.google.com/scholar/about.html/

40. P. Jacso, Google Scholar Beta. Péter's Digital Reference Shelf, 2004, Retrieved Jan 10, 2006, from http://snipurl.com/dwco/

41. P. Jacso, Google Scholar: the pros and the cons. Online Information Review, 29 (2), 208-214, 2005.

42. G. R. Notess, Scholarly Web Searching: Google Scholar and Scirus. Online, 29(4), 2005.

43. R. Kling, and E. Callahan, Electronic journals, the internet, and scholarly publishing. Annual Review of Information Science and Technology, 37, 127-177, 2003.

44. J. Bar-Ilan, The use of Web search engines in information science research. Annual Review of Information Science and Technology, 38, 231-288, 2004.

45. L. Vaughan, New measurements for search engine evaluation proposed and tested. Information Processing & Management, 40(4), 677-691, 2004.

46. S. Lawrence, and C. L. Giles, Accessibility of information on the web. Nature, 400, 107-109, 1999.

47. J. Bar-Ilan, Search engine results over time - a case study on search engine stability. Cybermetrics 2/3, 1999, Retrieved January 26, 2006, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

48. W. Mettrop and P. Nieuwenhuysen, Internet search engines - fluctuations in document accessibility. Journal of Documentation, 57(5), 623-651, 2001.

49. P. Mayr and F. Tosques, Google Web APIs: An instrument for webometric analyses? 2005, Retrieved January 20, 2006, from http://www.ib.hu-berlin.de/%7Emayr/arbeiten/ISSI2005_Mayr_Toques.pdf

50. J. Fry, and S. Talja, 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 (Vol. 41, pp. 20-30): Medford, NJ.: Information Today.

51. R. Whitley, The intellectual and social organization of the sciences (2 ed.). Oxford: Oxford University Press, 2000.