Scholarly Use of the Web:

What are the Key Inducers of Links to Journal Web Sites?

Liwen Vaughan[1]
Faculty of Information and Media Studies
University of Western Ontario
London, Ontario, N6A 5B7, Canada
Phone: (519) 661-2111 ext. 88499
Fax: (519) 661-3506
E-mail: lvaughan@uwo.ca
Mike Thelwall
School of Computing and Information Technology, University of Wolverhampton,
35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK.
Phone: + 44 1902 321470
Fax: + 44 1902 321478
E-mail: m.thelwall@wlv.ac.uk

Abstract

Web links have been studied by information scientists for at least six years but it is only in the past two that clear evidence has emerged to show that counts of links to scholarly Web spaces (universities and departments) can correlate significantly with research measures, giving some credence to their use for the investigation of scholarly communication. This paper reports on a study to investigate the factors that influence the creation of links to journal Web sites. An empirical approach is used: collecting data and testing for significant patterns. The specific questions addressed are whether site age and site content are inducers of links to a journal's Web site as measured by the ratio of link counts to Journal Impact Factors, two variables previously discovered to be related. A new methodology for data collection is also introduced that uses the Internet Archive to obtain an earliest known creation date for Web sites. The results show that both site age and site content are significant factors for the disciplines studied: library and information science, and law. Comparisons between the two fields also show disciplinary differences in Web site characteristics. Scholars and publishers should be particularly aware that richer content on a journal's Web site tends to generate links and thus the traffic to the site.

Introduction

The Web as an information source is an object of special interest to information scientists, and one that demands much new research (Spink, 2002). Of perhaps particular interest is scholarly use of the Web and hyperlinks as a collective source of new information (Cronin, 2001). The study of Web links contains both promise and pitfalls, however. The promise is that their use is not limited to the mainstream of scholarly research in the way that bibliographical citations are and therefore that they will have the potential to reveal types of information that were previously inaccessible or difficult to quantify (Davenport & Cronin, 2000). Conversely, anyone who can create a Web page and post it on the Internet could link to any other page, unimpeded by any quality control similar to that of a scholarly journal,

[1] *This is a preprint of an article to be published in the Journal of the American Society for Information Science and Technology* © *copyright 2002 John Wiley & Sons, Inc.* http://www.interscience.wiley.com/

creating problems with the interpretation of link counts. Web pages are also difficult to count and can change or disappear over time. Moreover, it is not possible to study the whole Web whatever tools are used (Thelwall, 2002a). Despite this, there are practical reasons to believe that the conceptualization by Davenport and Cronin (2000) of Web links as purveyors of trust in their targets is a good one, provided that a sufficient level of aggregation is used and methodological safeguards are taken. The causes for optimism are twofold: a series of significant correlations between Web link metrics and other measures in different contexts (described below); and the success of the link popularity algorithm pioneered by Google (Brin & Page, 1998) and similar link-based approaches adopted by other major search engines such as AltaVista (2002) and NorthernLight (2002).

Journal web sites could play a critical role in scholarly communication for three reasons: the increasing use of the Web as an information source both inside and outside academia; the centrality of journals in disseminating scientific research; and the astonishing increasing of the number of journals available through the Web in the last two years, including both the new electronic journals and traditional print journals having online versions. This has lead to the creation of sophisticated Web sites by the major publishing houses and a number of debates about access to scientific research. In this context there is a need to develop models to help understand the role of journal Web sites. For example, is the visibility of a Web site related to factors such as the impact of the associated journal, academic discipline, age of the Web site, or information content provided on the site? One measure of the visibility of a Web site is the number of links that lead to the site because the more links to a site, the more chances the site will be visited and therefore the more potential impact the site will have within the scholarly community. Researchers may also wish to know how the impact of their research is conditioned by the publicly available information about it that is posted on the Web (Lawrence, 2001).

Related Studies

Web link research in information science started with Larson (1996), Rousseau (1997), and Almind and Ingwersen (1997), all using an analogy between citations and hyperlinks to trigger explorations into the new phenomenon. This was followed by further investigations from Ingwersen (1998) and Leydesdorff and Curran (2000) amongst others, some of which are described below. New techniques developed included the use of the functionality of the search engine AltaVista to execute Boolean commands to find pages in one Web space that link to another. This search facility gave researchers the ability to access in a matter of seconds and without charge summary statistics concerning the interlinking of hundreds of millions of Web pages. On the back of this, Ingwersen (1998) proposed a new measure, the Web Impact Factor, modeled on the Journal Impact Factor (Garfield, 1994). Initial experiments produced disappointing results, however. Problems were found with the reliability of search engines (Rousseau, 1998/9; Smith, 1999; Snyder & Rosenbaum, 1999; Thelwall, 2000) and attempts to establish that link counts were meaningfully related to other known variables foundered (Sandvik, 1999; Smith, 1999; Thomas & Willett, 2000; Thelwall, 2001a, Soualmia *et al*., 2002). This lead to speculation that perhaps Web link analysis would not live up to its early promise (Bar-Ilan, 2001; Björneborn & Ingwersen, 2001; Thelwall, 2001a). But then a series of positive results emerged: AltaVista, the main search engine for data collection, had become more stable (Thelwall, 2001b); link counts to universities could correlate with accepted research ratings when the data is appropriately analyzed (Thelwall, 2001c, 2002b; Smith & Thelwall, 2002); links to faculties could correlate with rankings (Chu *et al*., 2002); advanced models for Web link counting could dramatically reduce outliers in data (Thelwall 2002c).

Journal Web sites are of particular importance if scholarly use of the Web is to be modeled or understood, primarily due to their association with the key communication medium of journal articles. A very approximate analogy can also be made between citation analysis and link analysis in this context, and one logical question to ask is whether link-based metrics can produce results that correlate with existing measures of the impact of the associated journal. Smith (1999) studied journal Web site inlink pages (i.e. pages outside a given site that contain at least one link to any page inside the site in question) and found counts of these to be an unreliable indicator of impact. Harter and Ford (2000) found no significant correlation between journal impact factors and e-journal inlink page counts, concluding that they were measuring something different. Later studies have successfully demonstrated statistically significant results, however, perhaps due to the maturing of the Web over time. For example, Vaughan and Hysen (2002) discovered a correlation between journal Web site inlink page counts and the journal's impact factor for the library and information science (LIS) field and the same association was subsequently revealed for law (Vaughan & Thelwall, 2002).

At the same time there has been surprisingly little research directly into the general causes of Web link creation, although some related findings have emerged. Links have been studied in electronic journal articles by interviewing authors (Kim, 2000) with the discovery of new reasons for linking that were not applicable to traditional citations. Goodrum *et al.* (2001) found that citation patterns for online scholarly articles differ from citation patterns found in the databases of the Institute of Scientific Information (ISI) because there are more online conference articles and these tend to cite other conference articles more than journal articles do. Kling and McKim (1999, 2000) have discussed extensively the issue of electronic publication and have concluded that heterogeneity will be an enduring feature of scholarly use of electronic communication media. However, there have been few empirical studies to test hypotheses such as those that arise from social informatics.

Research Questions

The association found between traditional impact measures and journal Web site inlink page counts is a reassuring indicator that the study of hyperlinks to these scholarly Web spaces is likely to reveal information about scholarly communication. The study reported here is an attempt to deepen the understanding of the factors involved in linking to journal Web sites in the belief that Web sites are an increasingly important component of research communication. It is also believed that the more links to the journal Web site, the more links and thus the more traffic to the site. The underlying model is that site age, site content and the discipline of the journal will affect the total inlink page counts. Other factors may well have a significant impact, for example site quality, but this would be very difficult to measure for empirical testing purposes. Accordingly, the primary hypotheses to be tested in the study are as follows.

- Links to journal Web sites (in the form of inlink page counts) will correlate positively with site ages.
- Links to journal Web sites (in the form of inlink page counts) will correlate with the type of information content on the site.
- There will be disciplinary differences in the Web link metrics which contrasts to that of disciplinary differences in the Journal Impact Factors.

The two disciplines chosen for a disciplinary comparison are library and information science (LIS) and law. The two disciplines are similar in that both are professional fields and have strong tradition of applied research. However, the two fields are predominantly non-overlapping with only one journal in common (the Law Library Journal). Furthermore, the use of the Web is an integral part of practice for information professionals while the same

cannot be said of the law, which provides an interesting contrast. It could be argued that two other disciplines with a stronger contrast (or the opposite, two with a closer tie) could be chosen for the study. Ultimately, different academic fields are related in different ways and there is no optimal pair of fields that would give an ideal comparison. Comparing different pairs of fields might reveal different information and the choice of specific fields is not particularly important for the purpose of knowing that discipline is a factor.

Research Design

Online Tools for Data Collection

The WayBack Machine and the Internet Archive

It is difficult to ascertain the age of a journal Web site. From the individual pages "last modified by" dates can be obtained, but these will not be helpful unless they were known to remain unchanged since their original creation date. An alternative strategy is to email the editor or Webmaster, but in the era of Spam, this is unlikely to achieve a high return rate and results may be biased by ephemeral factors such as newer sites' architects being more willing to communicate about their creation. One newly available alternative is the WayBack machine of the Internet Archive (www.archive.org). This is a very ambitious project that maintains a historical archive of data produced through crawls of the Web by a commercial search engine and other sources (Koman, 2002). See also Chavez-Demoulin *et al*. (2000) for a survey of the issues involved in this kind of venture. The WayBack machine is the name for a Web interface onto the database that gives a report of each time an URL has been crawled, with the additional provision of a link to the archived copy of the pages fetched. This information is highly reliable in the sense that the page retrieved in the archive can be viewed and there is no reason to believe that the crawl dates reported are anything but precise. This can be used to get an earliest known creation time for an URL. Some further points need to be made about this archive, however.

1. Search engines do not cover the whole Web (Lawrence & Giles, 1999) and so the absence of a Web page from the archive does not infer that it was not already created, only that the URL has not been discovered or indexed yet by the crawlers feeding the archive.
2. The URL is sensitive to the exact file name and path used. Many pages have multiple equivalent URLs, e.g. http://www.blackwellpublishers.co.uk/journals/JOLS/descript.htm and http://www.blackwellpublishers.co.uk/journals/JOLS/, but the archive records these URLs separately even if the pages are identical. In order to get a more reliable earliest known creation date, all possible versions of the URL must be checked separately.
3. Web sites and pages can ban robots from indexing them so that entire sites can be omitted. This has been the case for the publisher Elsevier(http://www.elsevier.nl/robots.txt, accessed on May 30, 2002).

The Commercial Search Engine AltaVista

The WayBack machine was not used for link counts because it did not provide this function. AltaVista was employed instead, but since it has been used for this purpose extensively before it will be only briefly discussed here. It has the generic search engine problems as the Internet Archive (1 and 3 above) and additional ones related to the specific task for which it will be used. AltaVista's advanced search permits Boolean requests that find all pages outside a given Web site that contain a link to it. Its reporting of results, however, is

subject to fluctuations principally because it will only search a fraction of its index and then use a probabilistic technique to estimate the number of matching pages in the whole database, according to its senior scientist (Broder, 2001). Since the amount of time spent on processing a query is dependant upon how busy the servers are, different search results can be returned at different times. This problem can be alleviated by conducting searches at the Web's low traffic times. We adopted this strategy for data collection and the search results were found to be very stable. Queries can also be 'remembered' (cached) instead of re-executed, which can give a false impression of stability. The study of Thelwall (2001b) did, however, find its results to be very stable over a period of seven months, but the figures should nevertheless be treated with caution and not regarded as in any sense a definitive and accurate count of "all" Web pages.

Data Collection

Locating Journal Web Sites

Journals in two fields, law and LIS, that were listed in ISI database were candidates for the study. Journal Impact Factors were obtained from the ISI Web of Science for the year 2000 (year 2001 data were not available at the time of the study). For each journal indexed by the ISI, the 2000 JIF is calculated to be the number of citations from ISI indexed journals published in 2000 to articles published in the years 1998 and 1999 by the journal in question, divided by the number of citable items published by the journal in the years 1998 and 1999 (Garfield, 1994).

URLs for these journal Web sites were sought through (a) searches for their titles as phrase queries in Google, (b) checking all known online lists of links to journals from the relevant discipline(s), and (c) checking for the official URL listed in the print copy of the journal. Multiple Web sites were found for some journals. For the purpose of this study, only the official Web site (e.g. journal publisher's Web site) was used. When multiple URLs were found for a site (e.g. different logical URL for the same physical site), all URLs were used in data collection and the results aggregated as detailed later.

There are 101 law journals and 53 LIS journals in year 2000 ISI database. However, some journals had to be excluded from the study for various reasons. For example, some journals do not have an independent Web site. Some journals have an URL that represents a database search query (which typically has a "?" in the middle of the URL). Search engines have difficulties finding links for this kind of URL. As a result, 38 LIS journals and 88 law journals were used for the data collection. See Appendix 1 and 2 for a list of the journals.

Web Site Age

The following steps were taken to identify an earliest known existence date for each journal's affiliated Web site.

1. Each directory associated with a discovered URL was visited to check whether it was exclusive to the journal or also contained other information. In many cases this was clear from the URL itself, e.g. www.jamia.org for JAMIA but in others this involved extensive browsing of the site. Each directory found to be exclusive to the journal was recorded.
2. Each URL found was checked in the Internet Archive and its earliest crawl date was recorded.
3. For each directory identified as belonging exclusively to the journal, eight different URLs were queried in the Archive: the full URL (with terminating slash); the URL without terminating slash; the full URL with the standard default file names (index.htm, index.html, home.htm, home.html, default.htm, default.html). The oldest of all these

results was then recorded. On Web servers, it is standard practice to have a file name associated with each directory root, and so it is helpful to attempt to identify that filename, because of its potential inclusion in the Archive. The procedure above attempts to do this by trying the most likely names. It is not 100% effective, however, since the Webmaster can choose almost any name to be the default, simply by changing the server initialization file.

4. For each journal the date recorded was the earliest one found in steps 2 and 3. This represents the earliest date when the Archive's crawler 'knew' any page about it to be on the Web. For convenience of the data analysis, the retrieved date was converted into the number of months before January 1, 2002.
5. Each earliest date found was double-checked to ensure that the earliest page retrieved was genuinely associated with the journal. For example, a search in the Archive for the URL http://www.jamia.org/ gave the earliest crawl date as Jan 15, 2000. This page was viewed from the Archive by clicking on the first link of the retrieved list. However, the archived copy concerned "Jamia Millia Islamia" rather than the journal sought. The next indexed date for the URL was March 8, 2000, and the archived page was for "The Journal of the American Medical Informatics Association". From this it could safely be inferred that the JAMIA site was created not later than March 8, 2000.

Although it was believed that the results from the Archive would be reliable in the sense of reproducible over time without changes, the data collection (steps 2 to 4) was rerun after an eleven-day period to see whether any changes had occurred in the results. Two changes were found. One site had recently posted a robots.txt file banning the crawler from visiting it. This resulted in the Archive not giving out any information at all about it. One site of a publisher had removed its robots.txt file, allowing the Archive spider to visit it. Two of the journals affected then showed a date from 1998 when they had presumably been crawled before the robots.txt file had been posted. Other than these changes, the two rounds of data collected are identical giving some confidence that the Internet Archive algorithm for checking its databases was reliable.

Web Sites Inlink Page Counts

For the purpose of extracting inlink page counts from AltaVista (the number of pages outside a given site that contain at least one link to any page inside the site in question), the following procedure was followed.

1. For each identified URL associated with the journal, if it was the only page for the journal on the site then the full URL was used. If not, then the shortest truncated URL uniquely identifying the journal was found. This needs a further explanation, as the need to truncate along 'word' boundaries is an undocumented feature of AltaVista. Each 'word' is a set of consecutive letters and numbers terminated by a non-alphanumeric character such as a dot, slash, minus sign, ampersand etc. For example, the URL http://www.jamia.org/misc/terms.shtml would be made up of the words www, jamia, org, misc, terms, shtml. Pages that link to this URL can be found by the command link:www.jamia.org/misc/terms.shtml, whereas pages that link to all pages with the same file name but any file extension can be found with link:www.jamia.org/misc/terms. However, link:www.jamia.org/misc/term will not work because the missing 's' at the end means that the word 'term' will be matched instead of 'terms'.
2. For each distinct partial URL obtained in step 1 the associated URL site identifying whole word partial domain name was identified and the following query executed. link:partial_URL AND NOT host:partial_domain. The partial domain is the righthand-most part of the domain name that identifies the site. For example in the case of www.jamia.org the partial domain would be jamia.org, in case there were other

subdomains that contained links to the journal, such as mail.jamia.org or backissues.jamia.org. In this example the query would be link:www.jamia.org AND NOT host:jamia.org, which should capture all pages with links to any page of the journal site, other than those from pages directly associated with it.

3. For each journal the inlink page count is the *total* of counts for each associated URL obtained in step 2.

The methodology described here is slightly different from that used in the previous paper (Vaughan & Thelwall, 2002) in that in step 2 partial domains are used instead of full domains in the "not host" part of the query. This is an improvement due to the exclusion of extra potential inlink pages from domains closely affiliated to the one hosting the journal.

The data collection was repeated after a period of 12 days and the average of the two scores used. Only one inlink page count had changed during this period and only by two pages. In the light of the surprising lack of variation, the two sets of AltaVista search result pages were compared and verified to be genuinely different, as a safeguard against caching. It was found that the second set contained additional Valentine's day text links, affirming that the pages themselves had not been retrieved from a cache.

Content Classification

The content of each journal Web site was classified into four groups as shown in Table 1. The classification was based on the hypothesis that more links would be attracted to a site with more journal content information. Table 1 is a hierarchical arrangement in that each category encapsulates the previous one.

TABLE 1. Categories for classifying Web site content.

Category	Description
Basic	The Web site contains basic information about the journal such as subscription information, editorial board members, and instructions to authors.
Title	The site provides free access to current table of contents. The table must cover at least one year.
Abstract	The site provides free access to a current list of abstracts. The list must cover at least one year.
Full-text	The site provides free access to a current set of full-text articles. The set must cover at least one year.

The reason for the stipulation of "at least one year" in the above classification scheme was the assumption that a reasonable period of time would be needed for any changes in content level to have an impact on the number of links to the site. It should be noted that this classification scheme does not represent an exhaustive list of all possible situations. A few journals did not fit neatly into this classification scheme as the following examples illustrate.

- Only the current issue of the journal is placed online.
- The journal has full-text online but it is password protected.
- Detailed information about the journal is embedded in other online databases but not the journal Web page itself. For example the DBLP server (http://dblp.uni-trier.de/) maintains a title list of Information Processing & Management.

In all cases, the journal Web sites were classified by ignoring this kind of additional partial information. The above cases were not incorporated into the classification scheme because of the need for a simple enough model for data analysis, although it is reasonable to assume that all would alter linking patterns to some extent. Their omission does not invalidate the results but may make it harder to get statistically significant conclusions.

Results

The data was first analyzed in order to test for the hypothesized key factors that induce links to journal Web sites. The data from the two disciplines was then compared to determine similarities and differences and the results are presented as three comparisons below.

Key Factor 1: The Impact of the Journal Within the Scholarly Community

The widely used journal impact factor (JIF) published by the ISI was employed as a measure of the impact of a journal within the scholarly community. It is recognized that JIF is not a perfect measurement for individual journals, but it is nevertheless contended that it is appropriate to use it within a discipline as an approximate impact measure because precision at the level of individual journals is not necessary for its use here (Cole, 2000; Adam, 2002; Moed, 2002). Although the JIF is normalized for journal size by dividing citations by citable items, no equivalent normalization was performed on inlink page counts. This follows from the hypothesis that inlinks predominantly target the journal Web site as an entity, typically via its home page, rather than individual pages (or articles, when available). For example, Harter and Ford (2000) found a median of 26 inlink pages linking to articles but a median of 135 for the whole site using a selection of e-journal Web sites containing online refereed articles. They compared total citations with total inlink pages, but conjectured that the two were measuring something different.

Spearman correlation coefficient tests were used to determine the relationship between the JIF and inlink page counts. The Spearman rather than the Pearson correlation coefficient test was used because the frequency distributions of the JIF and the inlink page counts were very skewed, particularly the latter. The test results showed a significant correlation ($p<0.01$) between the two variables for both law and LIS (correlation coefficients of 0.33 and 0.48 respectively). The same results were obtained in an earlier study using similar data sets (see Vaughan & Thelwall, 2002, for more detailed discussion on the relationship between the two variables). This means that journals having more (JIF) impact within the scholarly community attract more inlink pages.

Key Factor 2: Journal Web Site Content Level

Since the JIF has been shown to affect inlink page counts, we need to control its effect when examining the effect of content level on inlink page counts. This is achieved by developing a new measure, the link-JIF ratio, which is the inlink page count divided by the JIF score. The link-JIF ratio was calculated for each journal. If the JIF were the only factor that affected inlink page counts, then the link-JIF ratio would be the same for all journals. The different link-JIF ratio scores among journals allowed the effects of other factors to be examined while controlling for the JIF effect.

Each journal Web site in the study was classified into one of the four content groups: basic, title, abstract, and full-text as discussed in the methodology section of the paper. To determine if content level affects inlink page counts, a Kruskal-Wallis test was conducted which compared the link-JIF ratio of the four groups. The Kruskal-Wallis test was used, rather than a one-way analysis of variance test, because the frequency distribution of the inlink page count was very skewed (Howell, 2002, p. 323; Vaughan, 2001, p. 137). The Kruskal-Wallis test was applied to the LIS and law data separately. Both tests showed a significant difference ($p<0.01$) among the four content groups in their link-JIF ratios. Table 2 lists the link-JIF ratios for the four content groups, which shows a clear pattern that Web sites that have more content attracted more inlinking pages relative to their JIF scores with the exception of the title and abstract groups for LIS where the numbers are very close.

TABLE 2. Median Link-JIF ratio for the four Content Groups

Content Group	LIS	Law
Basic	31.65	7.24
Title	130.19	27.75
Abstract	103.95	32.75
Full-text	637.29	69.87

Key Factor 3: Journal Web Site Age

Spearman correlation coefficient tests were used to determine if the age of a journal Web site was related to its link-JIF ratio. The correlation coefficients were 0.468 and 0.436 for law and LIS respectively. The former is statistically significant at the 0.01 level and the latter at the 0.05 level. This means that older journal Web sites attracted more inlink pages.

It is important to determine whether the age of a Web site is related to the site's content level. If they were related, then the relationship between age and the link-JIF ratio identified here could be attributed to the relationship between content level and the link-JIF ratio, the second key factor discussed above. A one-way analysis of variance test was conducted with age as the dependent variable and content level as the independent variable. The test result shows insufficient evidence that content level affects age ($p=0.15$). Although there is possibly a very weak but not statistically significant association between age and content level, this is not enough to explain the significant separate association between both of them and link-JIF ratios. By using the link-JIF ratio rather than inlink page counts in analyzing factors age and content level, we have isolated the first factor from the remaining two. The analysis of variance test has subsequently isolated the age factor from the content level factor.

Comparison 1: Link-JIF Ratio

Inlink page counts among journal Web sites vary greatly for both disciplines. The median inlink page counts for law and LIS are 23 and 49 respectively. Although the latter figure is higher than the former, a Mann-Whitney test showed that this difference in inlink page count is not statistically significant ($p=0.14$). However, a Mann-Whitney test comparing the link-JIF ratio between the two disciplines showed a highly significant difference ($p<0.01$). It should be noted that we used the Mann-Whitney test instead of the independent t-test because the frequency distributions for inlink page counts and the link-JIF ratios are both very skewed (Howell, 2002, p. 707; Vaughan, 2001, p. 122).

The median link-JIF ratio is 19.18 and 99.54 for law and LIS respectively. This shows that LIS journal Web sites have more inlinking pages relative to their JIF scores. A possible explanation for this phenomenon is that information professionals in LIS make heavier use of the Web, which is a very important information medium. Additionally, legal professionals have stronger tradition in citation and law journals have higher JIF scores. A cursory look at ISI database reveals that average JIF score for law is among the highest of all disciplines.

Comparison 2: Content Level

When data from the two disciplines are combined, a minority of journal Web sites (44%) provide abstract or full-text access of its papers while more provide only basic or title information (56%). When the number of Web sites for each content group is tabulated separately for the two disciplines (see the first numbers in each cell of Table 3), law journal Web sites fall proportionally more into the basic and title groups. A chi-square test proved that there is a significant relationship ($p<0.05$) between content level and the discipline. Comparing the observed count (first number in each cell of Table 3) with the expected count (second number in each cell), law is over represented in the title category while LIS is over

represented in the full-text category. Overall LIS journal Web sites provide more content (55% provide abstract or full-text) than law journal Web sites (39% provide abstract or full-text).

TABLE 3. Content Level Comparison between Law and LIS

	Basic	Title	Abstract	Full-text
Law (observed count)	27	26	24	10
(expected count)	28.5	20.2	25.8	12.5
LIS (observed count)	14	3	13	8
(expected count)	12.5	8.8	11.2	5.5

Comparison 3: Journal Web Site Age

Data for Web site ages follow a normal distribution. This is true when the data from the two disciplines are combined or separated. Calculated as the number of months that the Web site has existed in January 2002, the average age for the law and LIS journal Web sites is 32.5 and 32 respectively. An independent t-test showed no significant difference ($p=0.66$) between the two disciplines in Web site ages. Translating the average age of 32 months into years, we can say that journal Web sites started at around April of 1999 on average.

Even the variability of the recorded Web site age is very similar between the two disciplines. The standard deviation is 16.4 and 15.8 respectively for law and LIS. Given that the average age is around 32, a standard deviation of around 16 is fairly large. Indeed, this is a wide range of about 56 months for both disciplines. The oldest Web sites started in October 1996 while the youngest began in July 2001.

Discussion

The study has introduced a new methodology that uses the Internet Archive for collection of date-related information. Various issues concerning the use of this new resource were explored and technical strategies for obtaining useful results were detailed in the methodology section of the paper. Issues surrounding the effective use of the commercial search engine AltaVista for data collection were also discussed. Three key factors inducing links to a journal Web site were identified: the impact of the journal, the content level of the site, and the age of the site. Firstly, it was confirmed that higher impact journals, as indicated by higher journal impact factor scores, attracted more inlink pages. Although the positive findings concerning link page counts apply only to the portion of the Web covered by AltaVista, this search engine does seem to have particularly good coverage of university Web sites (Thelwall, 2002c) and it seems reasonable to suppose that its coverage will not be systematically biased in favor of pages that would link to a particular class of journals. This first result serves as a partial assurance that the necessary restriction of data collection to the area covered by a search engine does not invalidate the conclusions reached.

The journal Web sites investigated were found to contain various level of content information with some providing only a basic description of the journal whilst others listed titles or abstracts or provided free access to full-text articles. The comparisons showed that Web sites which contained more content attracted more inlink pages. The connection between content and inlink pages provides evidence of the importance of the Web for scholarly communication: clearly the journals that are placing more information online are attracting more attention. The evidence here suggests that informative Web sites are valuable for journals, not just in the technological field of LIS but also in areas such as law for which technology is not such an integral part of the discipline. In the light of these findings, it is a logical possibility that journals which have been online longer and with more content have increased their visibility sufficiently to be reflected in their JIF. This has not been tested for,

however, and would need a much more complex research design to approach the problem of identifying one factor as causative of the other.

The age of a Web site also affects the number of inlink pages, with older ones receiving more. Caution must be exercised when interpreting this finding, however. In general, newly created links (e.g. links in the newest pages on the Web or new links that have been recently added to old pages) are less likely to have been indexed by AltaVista. This may introduce a confounding variable into the study. However, given the relatively frequent crawling undertaken by modern search engines and the fact that the youngest Web site in the study is six month old at the time of data collection, this should not significantly affect the conclusion reached. One explanation for this correlation is that it takes an appreciable time for Web page authors to identify and then create links to a journal Web site after its appearance. An alternative explanation is that Web authors infrequently or never update some pages once they have been created. The former model would cover situations such as library maintained Web resource lists if they were updated annually, whereas the latter would include permanent publication types such as online articles in free access full text e-journals.

While the three factors discussed above apply generally to both disciplines investigated in this study, law and LIS, comparisons were made between the two fields to determine similarities and differences. It was found that LIS journal Web sites attracted more inlink pages than their law counterparts that have the same journal impact factors. It may be that information professionals in the LIS field make more use of the Web. This may also be reflected in the fact that LIS journal Web sites also contained more content: while the majority (55%) of LIS journals provided abstract or full-text in their Web site, only 39% of the law journals did so. There was no difference between law and LIS in Web site age, however. The average identified starting point of Web sites was around April 1999 and both fields had a large variability (standard deviation 16 months) in Web site age.

One issue remains when interpreting the results of the statistical tests: that little is known about the origin of the link pages being counted. The impression gained during the research was that except for Internet World, which is more of a professional magazine than a scholarly journal, they are normally scholarly sources but they varied from resource lists compiled by librarians to student reading lists, general research links and own publication links created by academics. There will probably also be links created by students and practitioners, making the source much more heterogeneous than for the journal articles used for citation analysis. In the longer term, this could serve as a new source of information about the wider use and dissemination of scholarly information (Cronin, 2001), but it does complicate the problem of interpreting the inlink page counts. In this context, the association between inlink page counts and JIF scores is a reassuring indicator that links to journal Web sites are of a scholarly nature to some extent. It also serves to confirm that the data collected from commercial search engines, whilst not perfect, are sufficiently reliable to be able to produce statistically significant results. Based upon this success in using free Web resources for data collection, it is contended that link analysis is a promising area where more research remains to be done and more information on the patterns of Web use can be uncovered.

Conclusion

Evidence has been found to indicate that journal Web sites with more content are more visible in that they attract more links and therefore potentially more traffic to the sites. So publishers are encouraged to provide as much information as possible on their Web sites. Similarly, scholars may consider Web site content as a factor when choosing a journal to submit to, even in the case of print journals. Web site age has been show to affect site visibility: older Web sites are more visible. It could be reasoned that changes of URL are not

desirable because they can have a negative effect on the Web site visibility and thus reduce visits to the site. It also suggests that site age is a complicating factor that should be taken in to consideration when using link counts to compare Web sites. For the purposes of further research when comparing the impact of Web sites, journal citation impact and scholarly discipline must be taken into account, as these are both influencing factors. Moreover, although site inlink page counts do measure impact in a way that associates with average citation impact, it was also shown that the two are significantly different between disciplines. This suggests that they are actually measuring something different and therefore could be used in complimentary ways. The citation may measure the impact within the immediate research community while Web link may measure the impact to a wider constituency including practitioners and students. Further research is needed to gain a better understand of the nature of Web links. The disciplinary difference found also shows different patterns of Web use by different communities.

Acknowledgements

The referees are warmly thanked for their helpful comments. Many thanks to Ms. Kathy Hysen for locating some LIS journal Web sites.

References

Adam, D. (2002). The counting house, Nature, 415, 726-729.

Almind, T. C. & Ingwersen, P. (1997). Informetric analysis on the World Wide Web: methodological approaches to webometrics. Journal of Documentation, 53(4), 404-426.

AltaVista (2002). AltaVista Advanced Search Tutorial - Link Popularity. Available; http://help.altavista.com/adv_search/ast_haw_popularity, accessed 25 February, 2002.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes - A review and analysis. Scientometrics, 50(1), 7-32.

Björneborn, L. & Ingwersen, P. (2001). Perspectives of webometrics. Scientometrics, 50(1), 65-82.

Brin, S. and Page, L. (1998), The Anatomy of a large scale hypertextual web search engine, Computer Networks and ISDN Systems, Vol. 30 Nos. 1-7, pp. 107-117 Available at http://citeseer.nj.nec.com/brin98anatomy.html

Broder, A. (2001). Personal communication.

Chavez-Demoulin, V.C., Roehrl, A.S.A., Roehrl, R.A., Weinberg, A., (2000): The WEB archives: A time machine in your pocket, Internet Archive Colloquium, San Francisco, March 2000, Available http://citeseer.nj.nec.com/chavez-demoulin99web.html.

Chu, H., He, S. & Thelwall, M. (2002, to appear). Library and Information Science Schools in Canada and USA: A Webometric Perspective. Journal of Education for Library and Information Science.

Cole, J. R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In: Cronin, B. & Atkins, H. B. (eds.). The web of knowledge: a festschrift in honor of Eugene Garfield. Metford, NJ: Information Today Inc. ASIS Monograph Series, 281-300.

Cronin, B. (2001). Bibliometrics and Beyond: Some thoughts on web-based citation analysis. Journal of Information Science, 27(1), 1-7.

Davenport, E. & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In: Cronin, B. & Atkins, H. B. (eds.). The web of knowledge: a festschrift in honor of Eugene Garfield. Metford, NJ: Information Today Inc. ASIS Monograph Series, 517-534.

Garfield, E. (1994). The impact factor, Current Contents, June 20. Available: http://www.isinet.com/isi/hot/essays/journalcitationreports/7.html

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

Harter, S. & Ford, C. (2000). Web-based Analysis of E-journal Impact: Approaches, Problems, and Issues, Journal of the American Society for Information Science, 51(13), 1159-76.

Howell D. (2002). Statistical Methods for Psychology, 5th ed., Pacific Grove, CA, U.S.A.: Duxbury.

Ingwersen, P. (1998). The calculation of Web Impact Factors. Journal of Documentation, 54(2), 236-243.

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

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.

Kling, R. & McKim, G. (2000). Not Just a Matter of Time: Field Differences in the Shaping of Electronic Media in Supporting Scientific Communication. Journal of the American Society for Information Science, 51(14), 1306-1320.

Koman, R. (2002). How the Wayback Machine Works, Available: http://www.oreillynet.com/pub/a/webservices/2002/01/18/brewster.html, accessed 7 February, 2002.

Larson, R. R. (1996). Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace. ASIS 96. http://sherlock.berkeley.edu/asis96/asis96.html (visited 4 August 2001).

Lawrence, S. L. (2001) Online or invisible? Nature, 411(6837) 521.

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the web. Nature, 400, 107-109.

Leydesdorff, L. & Curran, M., (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy, Cybermetrics, 4. Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

Moed, H. F. (2002) The impact-factors debate: the ISI's uses and limits, Nature, 415, 731-732.

NorthernLight (2002). Northern Light General Help: Webmaster FAQs. Available: http://www.northernlight.com/docs/gen_help_faq_webmaster.html#rank, accessed 25 February, 2002.

Rousseau, R., (1997). Sitations, an exploratory study, Cybermetrics, 1. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, R. (1998/99). Daily Time Series of Common Single Word Searches in Alta Vista and Northern Light, Cybermetrics, 2/3(1). Available at http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html.

Sandvik H. (1999). Health information and interaction on the internet: a survey of female urinary incontinence, British Medical Journal, 319(7201), 29-32.

Smith, A. G. (1999). A tale of two web spaces: comparing sites using Web Impact Factors, Journal of Documentation, 55(5), 577-592.

Smith, A. & Thelwall, M. (2002). Web Impact Factors for Australasian Universities, Scientometrics, 54(1-2), 363-380.

Snyder, H. & Rosenbaum, H. (1999). Can Search Engines be Used as Tools for Web-link Analysis? A Critical View, Journal of Documentation, 55(4), 375-84.

Soualmia, L.F., Darmoni, S.J. Le Duff, F., Douyère, M., & Thelwall, M. (2002, to appear). Web Impact Factor: a bibliometric criterion applied to medical informatics societies'

Web sites, Medical Informatics in Europe MIE2002 congress (to be held in Budapest, Hungary, August 25-29).
Spink, A. (2002). Introduction to the special issue on Web research, Journal of the American Society for Information Science and Technology, 53(2), 65-66.
Thelwall, M. (2000), Web Impact Factors and Search Engine Coverage, Journal of Documentation, 56(2), 185-189.
Thelwall, M. (2001a). Results from a Web Impact Factor crawler, Journal of Documentation, 57(2), 177-191.
Thelwall, M. (2001b). The responsiveness of Search Engine Indexes. Cybermetrics, 5(1). Available: http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html.
Thelwall, M. (2001c). Extracting Macroscopic Information from Web Links, Journal of the American Society for Information Science and Technology, 52(13), 1157-68.
Thelwall, M. (2002a). Methodologies for Crawler Based Web Surveys, Internet Research: Electronic Networking and Applications, 12(2), 124-138.
Thelwall, M. (2002b). A comparison of sources of Links for academic Web Impact Factor Calculations. Journal of Documentation, 58, 60-72.
Thelwall, M. (2002c, to appear) Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites, Journal of the American Society for Information Science and Technology.
Thomas, O. and Willett, P. (2000). Webometric Analysis of Departments of Librarianship and Information Science. Journal of Information Science, 26(6), 421-428.
Vaughan, L. (2001). Statistical Methods for the Information Professional: A Practical, Painless Approach to Understanding, Using, and Interpreting Statistics (ASIST Monograph Series), Medford, New Jersey: Information Today, Inc.
Vaughan, L. & Hysen, K. (2002, to appear). The Impact of Journal Websites, ASLIB Proceedings.
Vaughan, L. & Thelwall, M. (2002, to appear). Web link counts correlate with ISI Impact Factors: Evidence from two Disciplines, Proceedings of the Annual Conference of the American Society for Information Science and Technology.

Appendix 1: LIS Journals used in the study

LIS Journal Title	JIF	Inlink page counts
ASLIB Proceedings	0.397	91
Bulletin of the Medical Library Association	0.343	113
Canadian Journal of Information and Library Science	0.167	0
College & Research Libraries	0.905	20
Electronic Library	0.190	27
Government Information Quarterly	0.190	4
Information & Management	0.683	110
Information Processing & Management	0.719	185
Information Society	0.404	18
Information Systems Research	1.093	133
Information Technology and Libraries	0.481	231
Interlending & Document Supply	0.400	5
Internet World	1.167	18727
International Journal of Geographical Information Science	0.988	109
International Journal of Information Management	0.424	66
Journal of Academic Librarianship	0.296	13
Journal of the American Medical Informatics Association	3.089	1489
Journal of Documentation	1.640	206
Journal of Government Information	0.328	14
Journal of Health Communication	0.463	216
Journal of Information Science	0.473	14
Journal of Librarianship and Information Science	0.263	0
Journal of the American Society for Information Science and Technology	1.226	515
Law Library Journal	0.105	16
Library and Information Science Research	0.297	4
Library Journal	0.265	1089
Library Quarterly	0.407	184
Library Resources & Technical Services	0.361	64
Library Trends	0.316	6
Libri	0.188	0
MIS Quarterly	2.064	1959
NFD Information-Wissenschaft und Praxis	0.029	122
Online	0.456	3340
Program – Electronic Library and Information Systems	0.364	0
Restaurator – International Journal for the Preservation of Library and Archival Material	0.759	1
Scientometrics	0.660	0
Social Science Information	0.264	20
Telecommunications Policy	0.731	22

Appendix 2: Law Journals used in the study

Law Journal Title	JIF	Inlink page counts
Aba Journal	0.534	807
Administrative Law Review	0.6	147
American Bankruptcy Law Journal	1.294	4
American Business Law Journal	0.633	9
American Criminal Law Review	1.125	57
American Journal of Comparative Law	2.114	24
American Journal of International Law	2.667	246
American Journal of Law & Medicine	1.594	192
Boston University Law Review	1.254	6
Buffalo Law Review	1.043	443
Business Lawyer	0.802	0
California Law Review	2.695	83
Chinese Law and Government	0.033	5
Columbia Journal of Law and Social Problems	0.3	11
Columbia Journal of Transnational Law	1.139	20
Columbia Law Review	4.7	97
Common Market Law Review	1.367	35
Cornell International Law Journal	1	22
Cornell Law Review	3.517	163
Criminal Law Review	0.552	2
Denver University Law Review	0.141	50
Duke Law Journal	3.13	531
Ecology Law Quarterly	1	208
Employee Relations Law Journal	0.104	0
Family Law Quarterly	1.632	4
Food and Drug Law Journal	0.771	19
Fordham Law Review	1.309	28
George Washington Law Review	1.311	3
Georgetown Law Journal	3.353	72
Harvard Civil Rights-Civil Liberties Law Review	1.786	183
Harvard Environmental Law Review	2.217	228
Harvard International Law Journal	2.667	194
Harvard Journal of Law and Public Policy	1.019	2
Harvard Journal On Legislation	0.912	8
Harvard Law Review	6.347	300
Hastings Law Journal	1.04	246
IIC-International Review Of Industrial Property and Copyright Law	0.482	2
Indiana Law Journal	1.844	522
International Journal of Law and Psychiatry	0.662	41
International Journal of The Sociology Of Law	0.158	110
International Review Of Law and Economics	0.54	55
Iowa Law Review	1.381	10
Journal of Criminal Law & Criminology	1.15	1
Journal of Law & Economics	1.05	2008
Journal of Law and Society	0.691	5

Journal of Law Economics & Organization	1.383	239
Journal of Legal Education	0.646	10
Journal of Legal Medicine	0.733	17
Journal of Legal Studies	4.468	256
Journal of Maritime Law and Commerce	0.712	148
Journal of The American Academy Of Psychiatry and The Law	0.686	25
Journal of The Copyright Society Of The USA	0.947	7
Journal of World Trade	0.539	46
Judicature	0.431	20
Juvenile and Family Court Journal	0.279	0
Law & Society Review	1.778	10
Law and Human Behavior	1.861	45
Law and Social Inquiry-Journal of The American Bar Foundation	1.154	0
Law Library Journal	0.105	18
Louisiana Law Review	0.319	0
Michigan Law Review	4.646	24
Military Law Review	0.303	2
Minnesota Law Review	2.217	22
Natural Resources Journal	0.537	1
New York University Law Review	3.386	283
Northwestern University Law Review	3.278	4
Notre Dame Law Review	1.261	1
Ocean Development and International Law	0.458	43
Psychology Crime & Law	0.389	2
Psychology Public Policy and Law	2.407	111
Rutgers Law Review	0.571	16
Southern California Law Review	1.873	3
Stanford Journal of International Law	0.471	194
Stanford Law Review	5.414	86
Temple Law Review	1.294	6
UCLA Law Review	2.765	72
University Of Chicago Law Review	2.338	19
University Of Cincinnati Law Review	0.857	3
University Of Illinois Law Review	1.352	5
University Of Pennsylvania Journal of International Economic Law	0.667	150
University Of Pennsylvania Law Review	3.278	81
University Of Pittsburgh Law Review	0.919	4
Vanderbilt Law Review	2.718	2
Virginia Law Review	4.029	86
Washington Law Review	0.667	309
Washington Quarterly	0.167	344
Wisconsin Law Review	1.134	15
Yale Law Journal	4.729	386