Edited version appears in: *Nature*, Volume 411, Number 6837, p. 521, 2001.

Online or invisible?

Steve Lawrence
NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
lawrence@research.nj.nec.com

Abstract

Articles freely available online are more highly cited. For greater impact and faster scientific progress, authors and publishers should aim to make research easy to access.

The volume of scientific literature typically far exceeds the ability of scientists to identify and utilize all relevant information in their research. Improvements to the accessibility of scientific literature, allowing scientists to locate more relevant research within a given time, have the potential to dramatically improve communication and progress in science. With the web, scientists now have very convenient access to an increasing amount of literature that previously required trips to the library, inter-library loan delays, or substantial effort in locating the source. Evidence shows that usage increases when access is more convenient [2], and maximizing the usage of the scientific record benefits all of society.

Although availability varies greatly by discipline, over a million research articles are freely available on the web. Some journals and conferences provide free access online, others allow authors to post articles on the web, and others allow authors to purchase the right to post their articles on the web.

In this article we investigate the impact of free online availability by analyzing citation rates. We do not discuss methods of creating free online availability, such as time-delayed release or publication/membership/conference charges. Online availability of an article may not be expected to greatly improve access and impact by itself. For example, efficient means of locating articles via web search engines or specialized search services is required, and a substantial percentage of the literature needs to be indexed by these search services before it is worthwhile for many scientists to use them. Computer science is a forerunner in web availability – a substantial percentage of the literature is online and available through search engines such as Google (google.com), or specialized services such as ResearchIndex [1] (researchindex.org). Even so, the greatest impact of the online availability of computer science literature is likely yet to come, because comprehensive search services and more powerful search methods have only become available recently.

We analyzed 119,924 conference articles in computer science and related disciplines, obtained from DBLP (dblp.uni-trier.de). In computer science, conference articles are typically formal publications and are often more prestigious than journal articles, with acceptance rates at some conferences below 10%. Citation counts and online availability were estimated using ResearchIndex. The analysis excludes self-citations, where a citation is considered to be a self-citation if one or more of the citing and cited authors match.

Figure 1 shows the probability that an article is freely available online as a function of the number of citations to the article, and the year of publication of the article. The results are dramatic. There is a clear correlation between the number of times an article is cited, and the probability that the article is online. More highly cited articles, and more recent articles, are significantly more likely to be online.

The mean number of citations to offline articles is 2.74, and the mean number of citations to online articles is 7.03, or 2.6 times greater than the number for offline articles. These numbers mask variations over time – in particular, older articles have more citations on average, and older articles are less likely to be online. When considering articles within each year, and averaging across all years from 1990 to 2000, we find that online articles are cited 4.5 times more often than offline articles.

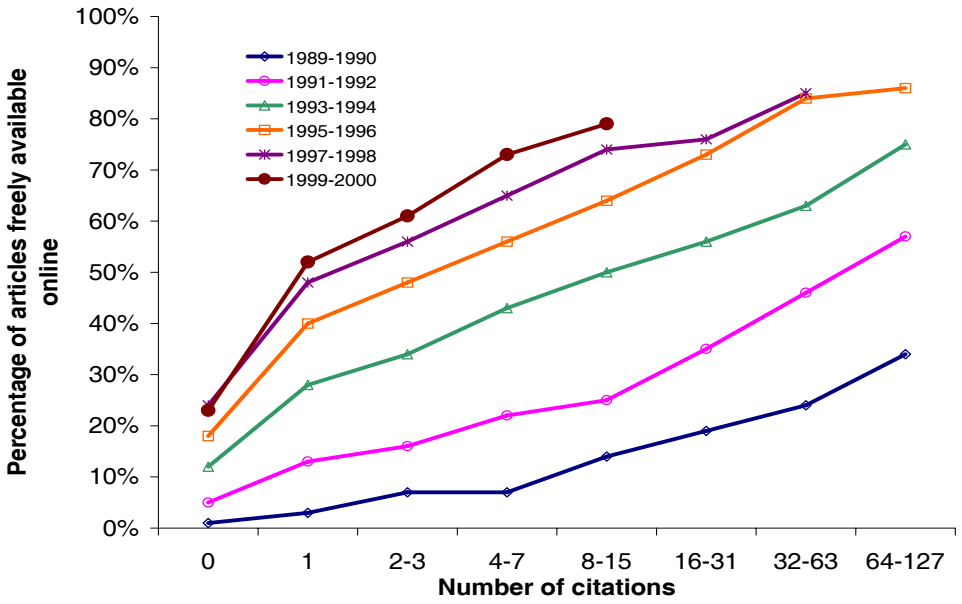

Figure 1. Analysis of 119,924 conference articles in computer science and related disciplines. More highly cited articles, and more recent articles, are substantially more likely to be freely available on the web. The actual percentage of articles available online is greater due to limitations in the extraction of article information from online documents, and limitations in locating articles on the web. Only points with greater than 100 articles are computed.

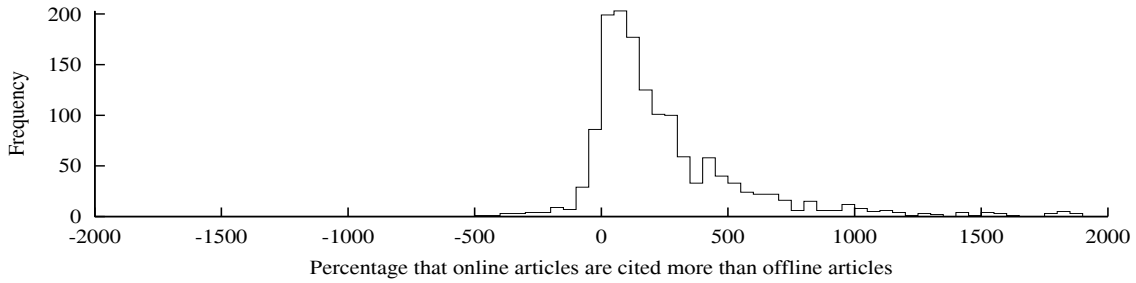

Figure 2. Analysis of citation rates within publication venues. The graph shows the distribution of the percentage increase for the average number of citations to online articles compared to offline articles. The analysis covers 1,494 publication venues containing at least 5 online and 5 offline articles. For 90% of venues, online articles are more highly cited on average. On average there are 336% more citations to online articles compared to offline articles published in the same venue [the first, second (median), and third quartiles of the distribution are 58%, 158%, and 361%].

We also analyzed differences within each publication venue, where multiple years for the same conference are considered as separate venues. We computed the percentage increase in the average number of citations to online articles compared to offline articles. When offline articles were more highly cited, we used the negative of the percentage increase for offline articles. For example, if the average number of citations for offline articles is 2, and the average for online articles is 4, the percentage increase would be 100%. For the opposite situation, the percentage increase would be -100%. Figure 2 shows the results. Averaging the percentage increase across 1,494 venues containing at least five offline and five online articles results in an average of 336% more citations to online articles compared to offline articles published in the same venue [the first, second (median), and third quartiles of the distribution are 58%, 158%, and 361%].

The preceding data does not allow us to make conclusions as to the cause of the correlation between high

citation rates and online availability. Online articles may be more highly cited because they are easier to access and thus more visible and more likely to be read, or because higher quality articles are more likely to be made available online. Intuitively, it seems likely that the easier availability and improved visibility of online articles plays a significant role. If we assume that articles published in the same venue are of similar quality, then the analysis by venue suggests that online articles are more highly cited because of their easier availability. This assumption is likely to be more valid for top-tier conferences with very high acceptance standards. Restricting the above analysis to the top publication venues by average citation rate results in a similarly dramatic increase in citation rates for online articles. For example, when restricting to the top 20 venues, the average increase in the citation rate for online articles is 286% [the first, second (median), and third quartiles of the distribution are 66%, 284%, and 471%].

Free online availability facilitates access in multiple ways, including online archives, direct connections between scientists or research groups, hassle-free links from email, discussion groups, and other services, indexing by web search engines, and the creation of third-party search services. Free online availability of scientific literature offers substantial benefits to science and society. To maximize impact, minimize redundancy, and speed scientific progress, author and publishers should aim to make research easy to access.

Acknowledgments

Thanks to Gary Flake, Andrew Odlyzko, and David Pennock for useful comments and suggestions.

References

[1] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[2] A. Odlyzko. The rapid evolution of scholarly communication. *Learned Publishing*, 2001. to appear.