2014

Decrease in Free Computer Science Papers Found through Google Scholar

Lee A. Pedersen
Brown University, Lee_Pedersen@brown.edu

Julie Arendt
Virginia Commonwealth University, jaarendt@vcu.edu

Recommended Citation

Lee A. Pedersen, Julie Arendt, (2014) "Decrease in free computer science papers found through Google Scholar", Online Information Review, Vol. 38 Iss: 3, pp.348 - 361.

Decrease in free computer science papers found through Google Scholar

Abstract:

Purpose - Google Scholar was used to locate free full-text versions of computer science research papers to determine what proportion could be freely accessed.

Design/methodology/approach - A sample of 1967 conference papers and periodical articles from 2003-2010, indexed in the ACM Guide to Computing Literature, was searched for manually in Google Scholar, using the paper or article title and the first author's surname and supplementary searches as needed.

Findings - Free full-text versions were found for 52% of the conference papers and 55% of the periodical articles. Documents with older publication dates were more likely to be freely accessible than newer documents, with free full-text versions found for 71% of items published in 2003 and 43% of items published 2010. Many documents did not indicate what version of the document was presented.

Research limitations/implications - Results were limited to the retrieval of known computer science publications via Google Scholar. The results may be different for other computer science publications, subject areas, types of searches, or search engines.

Practical implications - Users of Google Scholar for finding free full-text computer science research papers may be hindered by the lower access to recent publications. Because many papers are freely available, libraries and scholarly publishers may be better served by promoting services they provide beyond simple access to papers.

Originality/value – Previous research showed lower levels of free access than we found for computer science, but the decline found in this study runs contrary to increases found in previous research.

Keywords: Open access, Google Scholar, Computer science, Document versions

Article Classification: Research Paper

Introduction

As information professionals who have served computer science communities at universities, our users informed us that they could find research papers by searching with Google or, more recently, Google Scholar. However, they still expected the university to provide core computer science journals and conference proceedings. We wanted to determine how easy it was to access the content of computer science research publications by searching the web without subscription-based access through the library.

Others also have noted a trend of faculty using free online papers. Some computer science and engineering faculty interviewed at the College of New Jersey said that they used Google Scholar rather than Compendex because of the easier access to full text (Tucci, 2011). In a recent Ithaka survey, most faculty indicated “materials that are freely available online” were a very important source for scholarly publications and often used these free materials when their libraries did not have immediate access (Housewright, *et al.*, 2013, pp. 36-37). From 2003 to 2012, faculty increased their use of search engines as a starting point for their research, according to national surveys (Housewright *et al*., 2013).

Computer scientists were among the first to make full-text versions of their publications freely available electronically (Swan and Brown, 2005). Because of their early adoption of what is sometimes called self-archiving or green open access, computer science papers may be widely available for free. On the other hand, library subscriptions from publishers can allow on-campus users to access publications without any login or direct payment from the user, including subscription-based publications

found through Google Scholar. There is a wide gap between on-campus and off-campus access to publications found through Google Scholar (Christianson, 2007). Users may perceive easy access to free documents via a search engine, even though library subscriptions are paying for access. How much really is free? Has the free availability of computer science papers reached the point that it can supplant, rather than supplement, university library collections as the primary source of access?

Background

One recent multidisciplinary study used a software robot to crawl the Web for over one hundred thousand journal articles indexed in Web of Science in fourteen disciplines published from 2005 to 2010 and found an average open access percentage of 24% with a low of 10% for arts and a high of 45% for mathematics (Gargouri, 2012). In another multidisciplinary study, researchers used Google in fall 2009 to search for 1837 articles published in 2008 and found that 20.4% had open access availability (Björk *et al.*, 2010). Burns (2013) found 58% open access, but his study used a sample of papers recorded by readers in CiteULike rather than a random sample of papers published.

In the wide-ranging literature on open access, a few studies have investigated the availability of free access in computer science. Zhuang *et al*. (2005) programmed a focused web crawler to locate papers in computer science by starting from the authors' home pages. Their crawler was able to harvest over 81% of papers from *ACM SIGMOD International Workshop on the Web and Databases* and over 79% of papers from the

Journal of Artificial Intelligence Research. If this high success percentage were typical, freely available papers could fill the majority of a researcher's needs.

The high percentage of free full-text papers that Zhuang *et al*. (2005) found may not be typical of computer science papers. They chose a highly selective workshop and a highly selective journal (Zhuang *et al*., 2005), but research in other disciplines has found a relationship that heavily cited publications are also more likely to be freely available online, with their free availability possibly contributing to their higher citation counts (Wren, 2005; Moed, 2007). Other research on computer science publications showed lower availability for free full-text versions of papers. Silva *et al*. (2009) sent automated queries for computer science conference papers by Brazilian researchers to multiple search engines. Google Scholar was the best performing of the search engines, with a mean average precision of 32%, for locating a free copy of the full text of a paper. Silva *et al*. (2009) were able to improve this percentage to 40% by combining Google Scholar and Google results and to further improve the percentage to 45% by using an algorithm to re-rank the combined results set.

These studies only give a partial picture of how easy it is to get free full-text versions of computer science papers. Because the published studies of computer science publications rely on automated queries, their success percentages may not reflect that of human searchers. We instead, use naturalistic searching, as described by Christianson (2007), "By naturalistic is meant a setting in which humans, not machines, perform plausibly 'real-life' searches in Google Scholar." We demonstrate a naturalistic success percentage of using Google Scholar without a library subscription to access the computer science research literature.

Methodology

We used the ACM Guide to Computing Literature (ACM Guide) to retrieve the bibliographic citations for a sample of 2003-2010 computer science conference papers and periodical articles. For each citation in our sample, we queried Google Scholar as our users might. For each item sought, we noted whether we could open a free version of it.

Source of Bibliographic Citations

The first step was to find a comprehensive source of metadata for computer science research publications. Many organizations publish or index computer science research literature. We chose the ACM Guide. The ACM Guide includes all publications from the ACM, publishers affiliated with ACM, and other computer science publishers. It has strong coverage of both computer science conference proceedings and journals. The ACM Guide provides clear boundaries for what counts as a computer science publication, is comprehensive, and is freely available. Inspec also has a clear computer science category and strong coverage, but it was not readily available to both authors. We considered other alternatives such as Web of Science, DBLP Computer Science Bibliography, and IEEE Xplore as well, but we selected the ACM Guide because of its overall strengths. As Hennessey (2012, p. 34) describes, the ACM Guide, "...has the best content of databases" in the field of computer science.

Sample selection

Because it was not feasible for us to search for all of the resources indexed in the ACM Guide, we used a sample. For this study, we restricted the years of publication to 2003-2010. We limited the sample to the years 2003-2010 to cover slightly more than the average age of articles that are cited in computer science research. In the computer science categories of the Journal Citation Index, the aggregate cited half-life, i.e., the median age of articles cited in those areas, is around seven years, with some categories having an average cited half-life closer to nine years (Thomson Reuters, 2013). Similarly, a study of references from computer science journals and conferences showed that they refer to articles and papers that, on average, are seven years old, with references to conferences tending toward newer literature and references to journals tending to older literature (Wainer *et al.*, 2011). Our eight-year range is a bit longer than the approximate seven-year average.

We included conference papers as well as articles in periodicals in the sample. Conference papers were included because they play an important role in scholarly communication in computer science (Franceschet, 2010), but we excluded other works such as books, technical reports, and dissertations.

We stratified our sample by year. For statistical power, we sampled a minimum of 130 records per year. The number of records sampled for a year was proportionate to the number of entries in the ACM Guide for that year at the time the sampling was done, as shown in Table I. Because the ACM Guide had more than three and a half times as many entries for 2009 as it had for 2003, our sample also had more than three and a half times as many items for 2009 as for 2003. The total size of the initial sample was

2107 records. We sampled records from the ACM Guide during the week of January 16-20, 2012. The ACM regularly adds records to the ACM Guide, including records for publications that are more than a year old. Therefore, the proportion of items from each year only reflects the proportions that were present at the time we set the sample frame.

INSERT Table I (SAMPLE SIZE PER YEAR) HERE

With permission from the ACM, we manually gathered the sample from the ACM Guide. To do this, we entered an empty space as the query term into the search box in the ACM Digital Library (http://dl.acm.org/dl.cfm), expanded the result to include all records in the ACM Guide, narrowed the results to a single year, narrowed the results to periodicals and proceedings, and sorted the results by date. By sorting by date, each item in the results list had a unique, consistent number next to it. Then we used the "randbetween" function in Google Spreadsheet to generate random numbers. The items in the results list corresponding to those random numbers were included in the sample. We copied the bibliographic information from the selected items into our data spreadsheet, as shown in Figure 4 in the Appendix. Based on the title of the publication that an item appeared in, we also recorded whether it was a conference paper or a periodical article.

With this selection procedure, the sample included some non-content items such as cover art, lists of reviewers, list of keynote speakers for a conference, and title pages. A discussion with ACM confirmed that such records, which were not relevant to this research, could not be automatically filtered out. These items from the sample were

excluded from this study and were not replaced in the sample. As a rule of thumb, items were excluded if they primarily dealt with the publication or conference organization process rather than computer science research. Items that could have computer science subject matter, such as editorials and historic reviews, were included. A list of the types of items included and excluded is shown in the Appendix. The initial and final numbers of items in the sample are shown in Table I.

Google Scholar search technique

We standardized our search methods to model an intelligent searcher who is reasonably persistent. The search emphasized locating a complete document regardless of whether it was the "version of record" (NISO, 2008). All searches were conducted from the authors' homes, not connected to any employer or library proxy or VPN service. Each author searched for half of the documents for each year of the sample.

We chose to limit the search to a single search engine. Based on Silva ***et al***.'s (2008) finding that Google Scholar outperformed other search engines for a similar task and based on our pretesting, we used Google Scholar.

Silva ***et al.*** (2008) also compared several sets of search terms for locating papers. They were most successful in locating conference papers in Google Scholar by entering the entire title of a paper, without quotes, and the surname of the first author, so we used this technique. In copying the information from the ACM Guide to the Google Scholar search box, we did not standardize the capitalization of the entries into the box.

As we searched “naturalistically” (Christianson, 2007), rather than using an algorithm, we limited our efforts for the sake of time. We limited the first level of search to the ten items on the first page of the Google Scholar results list, much like many searchers (Spink *et al.*, 2002; Granka *et al.*, 2004; Rieger, 2009). We looked through the first page of results for any entries that matched the bibliographic information we had.

Google Scholar presented each result item as a snippet with some bibliographic information, as shown in Figure 1. When we found a snippet that matched the bibliographic information, we did not click the title link for the item because that link typically led to a published version behind a paywall. Shortcuts to free full-text documents often were presented on the right of the screen, and an “All [number] versions” link was presented below the bibliographic information. Instead of clicking on the main link, we followed a two-part process to locate free full-text documents.

INSERT Figure 1 (SCREENSHOT OF GOOGLE SCHOLAR RESULT) HERE

First, when available, we tried the shortcuts to full text. When a shortcut to full text led to a free, complete document, we recorded this result as a “Shortcut Full Text”, copied the URL into our data spreadsheet, and stopped searching.

Second, if the entry did not have “Shortcut Full Text”, we used the “All [number] versions” link to see if any of these items provided a free full-text document. Each version link in a list like that in Figure 2 was checked sequentially for whether it led to a full-text document. If a full-text item was found, we recorded that as “Version Full Text”,

copied the URL into our data spreadsheet, and stopped searching. Unlike our examination of the initial results list, even when there were more than ten versions and the list spanned more than one page, we continued checking until we found a full-text document or until the list of versions was exhausted. Generally, the first link provided on the “versions” page was the main link provided on the results page. It, nevertheless, is possible that some free versions were missed because we did not use the main link on the results page directly.

INSERT Figure 2 (SCREENSHOT OF ALL [NUMBER] VERSIONS) HERE

If we found a snippet that matched the bibliographic information but neither the full-text shortcut on the right nor the Google Scholar versions connected to a free copy of a full-text document, we recorded the item as “No Free Copy” and stopped searching.

If the initial search method of entering the entire title of a paper, without quotes, and the surname of the first author yielded no match to the bibliographic information in the Google Scholar results, we refined the search strategy. In no specific order, we used several methods: entering the title with any diacritical marks, symbols, or punctuation removed; adding double quotes (exact phrase syntax); adding the surname of the second author; and adding the source periodical or conference title. Although we did not record the success of each of these methods, removing diacritical marks and adding a second author’s surname seemed to be the most successful of these methods. We recorded these searches as “More than One Search” in our data entry sheet. When these searches yielded a snippet that matched the bibliographic information, we

followed the two-part procedure above. If none of the supplementary searches produced a matching snippet, we recorded the result as “Not Found at All.”

The data were collected between March 26 and May 29, 2012. The format of the data collection spreadsheet is shown in the Appendix. Counts were tabulated using IBM SPSS Statistics Version 20 and graphed using Microsoft Excel 2010.

Criteria for free full text

Because we found a variety of items, we set criteria for what was considered free full text. Nearly all multi-page documents with the correct title and authors, as recorded in the ACM Guide, were treated as full-text versions. Even if the item found lacked page numbering, a publisher’s imprint, or other markers of authenticity, we counted it as free full text if the item looked like it was complete. A version found in Google Books was considered free full text if the entire paper was readable, with no missing pages. Postscript, PDF, and HTML formats were acceptable as free full text so long as we could view what appeared to be a complete document. Extended abstracts were included as free full text if page numbers given in the ACM Guide indicated that the original was similarly short, but brief, one-paragraph abstracts were not considered full text. Items that clearly were missing sections, such as those missing pages or obviously missing figures or tables, were not treated as free full text. We also excluded PowerPoint slides and files that could not be opened.

We wanted to record whether the papers we found were the final, published versions of the documents and whether the versions had undergone peer review. Some of the papers clearly indicated that they were the version of record or were so poorly

formatted that we could infer that they were not the version of record. Many papers, though, were not clearly one or the other. For example, some papers had a two-column layout and contained an ACM or other copyright notice but had Xs in place of digits for ISSN numbers. Because we did not believe that we could consistently determine the version we had, we abandoned our efforts to record the version.

Findings

The initial sample from the ACM Guide included 1339 bibliographic entries from conferences and 768 entries from journals. This ratio of entries from conferences and entries from journals was similar to the proportion in the ACM Guide. The data analysis excluded 140 non-content entries, leaving bibliographic entries for 1967 items in the final sample: 1289 conference papers and 678 periodical articles.

We could locate snippets in Google Scholar for nearly all of the items using our basic search procedure. The initial search strategy located snippets for all but 143 items. With the supplementary search strategies, we found an additional 108 items. Just 35 items were not found at all.

We found free versions of the papers for 1044 items (53%). Most, 898, of the free items could be opened directly from the shortcut on the results page, but 146 freely available items were accessed through the "Version Full Text." Free full-text items were found for 52% of the conference papers and 55% of the periodical articles across all years. This small difference was not statistically significant, (χ^2=1.12, df=1, p=0.29). As shown in Figure 3, date of publication was related to the proportion of free full-text items

found. The oldest items, from 2003, had the highest proportion of free full text at 71%. The newest items, from 2009 and 2010, had the lowest proportion of free full text, at 43%. Both periodical articles and conference papers had a decline, as shown in Figure 3.

INSERT Figure 3 (PERCENTAGE FREE FULL TEXT BY YEAR) HERE

Discussion

Percent of free full text

We were able to find free full text for slightly more than half of the computer science documents we sought. This percentage was lower than the approximately four-fifths free full text that Zhuang *et al*. (2005) found for a selective computer science journal and a selective computer science conference, but Silva et al. (2009) only had a mean average precision of 45% for a wider range of computer science conference papers.

The percentage of free full text that we found of greater than fifty percent for computer science was high compared to other disciplines. Recent multidisciplinary studies found an average open access percentage of around one fifth to one quarter (Gargouri, 2012; Björk *et al.*, 2010). Although Burns (2013) found 58% open access, his study used a sample of papers recorded by readers in CiteULike rather than a random sample of papers published.

Contrary to the findings of previous studies, we saw a decline in free availability of about 4% per year. Studies covering multiple disciplines have shown growth, rather

than a decline, in free availability of articles (Hajjem, *et al*., 2005; Gargouri *et al*., 2012; Kurata *et al.*, 2013). In computer science, one of the first studies on this topic found that recent papers were more likely to be freely available online than older papers (Lawrence, 2001).

The decline in free availability that we found has many possible explanations. It is possible that computer scientists are not posting their papers as often as they used to. It, however, is possible that the observed decrease came from delays rather than declines. Authors may delay posting or self-archiving papers following publication. Regardless of when a paper is posted, there is a delay between its appearance on the web and its discovery by Google's web crawlers. An additional delay, accidental or deliberate, may occur before the free full-text version appears in Google Scholar's search results. Further research, using surveys of authors or using other search engines, could clarify whether our result was particular to our searches in Google Scholar.

Document authenticity

The National Information Standards Organization (NISO) has developed recommended practices with standardized language, for describing the versions of a journal article that may appear online (NISO/ALPSP Journal Article Versions (JAV) Technical Working Group, 2008). Rarely did we find a document with a description of its version, let alone a document that used standardized language to describe it. When we found versions with this documentation, they usually were proofs and versions of record and occasionally were marked as accepted manuscripts.

Our difficulty in ascertaining document versions was not unique to this study. For example, Goodrum, *et al*. (2001) examined 500 randomly selected open access computer science articles; for 315 of those documents, there was not sufficient information on the document to identify even the type of source it was (book chapter, journal article, conference proceedings, technical report, etc.).

Limitations of the study

Although we used human searching, we did not replicate everyday searching. A typical searcher may not start with complete, correct citation information from the ACM Guide nor use a search strategy documented to be successful. Because we took this approach, we may have had greater success than a typical computer science searcher in Google Scholar. Our persistence in trying all of the versions that Google Scholar listed also may have led us to find a higher proportion of free full-text documents than a typical searcher would.

On the other hand, our selection of documents may have given us fewer free full-text documents than a typical searcher would find. In general, heavily cited papers are more likely to be open access than papers that are not heavily cited, and open access papers are more heavily cited than papers with restricted access (Wagner, 2010). We searched a random selection of articles and conference papers, but a computer science researcher may be interested in the most cited documents. We also could have found more items if we had expanded our search to other search engines beyond Google Scholar (Norris *et al.*, 2008; Silva, *et al*., 2009).

If we had searched for heavily cited documents rather than a random selection of documents, we also may have been more successful in ascertaining document versions. Goodrum, *et al*. (2001) were able to find complete citation information within highly cited, free computer science papers, but their random selection of free computer science papers was much less likely to contain this information. It is possible highly cited papers also would have better documentation of their version types than a random group.

Our results also are limited to a particular set of papers available during a particular time. For example, we accepted conference papers that were available in full text in Google Books as free full text. Google and conference proceedings publishers may negotiate or renegotiate what percentage of a proceeding is freely available to preview. Such negotiations could change the number of papers that are freely accessible in Google Books. In addition, we accepted any free full-text documents that we found. Other document versions also may not continue to be available over time. Some of the versions we found were complete copies of conference proceedings that, possibly illegally, were placed on sites unassociated with the conference. Other documents were located on the authors' personal websites. Such versions may appear as authors add papers to their sites or disappear as authors change employers or retire.

Implications for practice

Using Google Scholar, we were able to find free full-text versions of about half the computer science research papers that we sought. Depending on one's perspective, this half could be a metaphorical half-full glass or half-empty glass.

For computer scientists who do no not have easy access to library collections, the large accessibility of free full-text documents is appealing. That appeal is tempered, however, by lower availability of newer documents than of older documents. Computer science researchers could lack access to a substantial number of papers without subscriptions. These researchers may be better served by trying more than one search tool to find recent publications.

For computer scientists with easy access to library collections, Google Scholar can deliver more than half of the articles and conference papers. It connects to library subscriptions for a large portion of the papers they seek (Christianson, 2007). It connects to free, though often unauthenticated versions, of many papers that their libraries may not have.

Libraries are in a glass half full, half empty situation as well. This study showed that about half of computer science research publications cannot easily be accessed for free through Google Scholar, so libraries still have value as a resource provider. Conversely, the widespread free availability of papers means that access to full-text papers should not be the sole value that libraries provide to computer scientists.

For publishers, the easy availability of free publications can be a threat to their sales. Because of the limitations of these free full-text versions, however, researchers and their institutions still have reasons to pay for the version of record, so limitations in availability and authenticity should be attractive to publishers.

Because free versions of many computer science papers are widely available, libraries and scholarly publishers may be better served by promoting the value they provide through rapid access and the version of record rather than merely the access

they provide. Some of that value may not be adequate, though, if computer scientists believe the free versions are good enough. Even with those "good enough" unvetted versions, not everything is free. Paid access would be necessary to cover the large swath of the computer science literature that is not available for free.

References

Björk, B-C., Welling P., Laakso, M., Majlender, P., Hedlund, T., and Guðnason, G. (2010), "Open access to the scientific journal literature: situation 2009", *PLoS ONE*, Vol. 5 No. 6, e11273, available at: http://dx.doi.org/10.1371/journal.pone.0011273 (accessed 7 June 2013).

Burns, C.S. (2013), "Google Scholar and free or open access scholarly content: impact on academic libraries", poster presented at the 2013 Association for Library and Information Science Education (ALISE) Annual Conference, Doctoral Poster Session, 22-25 January, Seattle, Washington, USA, available at: http://hdl.handle.net/10355/16178 (accessed 6 June 2013).

Christianson, M. (2007), "Ecology articles in Google Scholar: levels of access to articles in core journals", *Issues in Science and Technology Librarianship*, No. 49, available at: http://dx.doi.org/10.5062/F4MS3QPD (accessed 7 June 2013)

Franceschet, M. (2010), "The role of conference publications in CS", *Communications of the ACM*, Vol. 53 No. 12, pp. 129-32.

Gargouri, Y., Larivière, V., Gingras, Y., Carr, L., and Harnad, S. (2012), "Green and gold open access percentages and growth, by discipline", paper presented at 17th

International Conference on Science and Technology Indicators (STI), 5-8 September, Montreal, Canada, available at: http://eprints.soton.ac.uk/340294/ (accessed 28 June 2013).

Goodrum, A.A., McCain, K.W., Lawrence, S., and Giles, C.L. (2001), "Scholarly publishing in the internet age: a citation analysis of computer science literature", *Information Processing & Management*, Vol. 37 No. 5, pp. 661-75.

Granka, L.A., Joachims, T., and Gay, G. (2004), "Eye-tracking analysis of user behavior in WWW search", In *SIGIR '04 Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM: New York, 478-9.

Hajjem, C., Harnad, S., and Gingras, Y. (2005), "Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, Vol. 28 No. 4, pp. 39-46, available at: http://sites.computer.org/debull/A05dec/hajjem.pdf (accessed 23 May 2013).

Hennessey, C.L. (2012), "ACM Digital Library", *The Charleston Advisor*, Vol. 13 No. 4, pp. 34-8.

Housewright, R., Schonfeld, R., and Wulfson, K. (2013), "Ithaka S+R US faculty survey 2012", available at: http://www.sr.ithaka.org/sites/default/files/reports/Ithaka_SR_US_Faculty_Survey_2012_FINAL.pdf (accessed 9 May 2013).

Kurata K., Morioka T., Yokoi K., and Matsubayashi M. (2013), "Remarkable growth of open access in the biomedical field: analysis of PubMed articles from 2006 to

2010", *PLoS ONE*, Vol. 8 No. 5, e60925, available at http://dx.doi.org/10.1371/journal.pone.0060925 (accessed 7 June 2013).

Lawrence, S. (2001), "Free online availability substantially increases a paper's impact", *Nature*, Vol. 411 No. 6837, p. 521.

Moed, H.F. (2007), "The effect of 'open access' on citation impact: an analysis of ArXiv's condensed matter section", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 2047–54.

NISO/ALPSP Journal Article Versions (JAV) Technical Working Group (2008), "Journal Article Versions (JAV): Recommendations of the NISO/ALPSP JAV Technical Working Group", recommended practice [NISO-RP-8-2008] available at: http://www.niso.org/publications/rp/RP-8-2008.pdf (accessed 7 June 2013).

Norris, M., Oppenheim, C., & Rowland, F. (2008), "Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR", *Online Information Review,* Vol. 32 No. 6, pp. 709-15.

Rieger, O. (2009), "Search engine use behavior of students and faculty: user perceptions and implications for future research", *First Monday*, Vol. 14 No. 12, available at: http://firstmonday.org/ojs/index.php/fm/article/view/2716/2385 (accessed 29 May 2013).

Silva, A.J.C., Gonçalvesa, M.A., Laender, A.H.F., Modesto, M.A.B., Cristo, M., and Ziviani, N. (2009), "Finding what is missing from a digital library: a case study in the computer science field", *Information Processing & Management*, Vol. 45 No. 3, pp. 380-91.

Spink, A., Jansen, B.J., Wolfram, D., and Saracevic, T. (2002), “From e-sex to e-commerce: web search changes”, *Computer*, Vol. 35 No. 3, 107-9.

Swan, A., and Brown, S. (2005), “Open access self-archiving: an author study”, available at: http://cogprints.org/4385/1/jisc2.pdf (accessed13 May 2013).

Thomson Reuters (2013), 2011 Journal Citation Reports® Science Edition, (accessed 29 April 2013).

Tucci, V.K. (2011), “Assessing information-seeking behavior of computer science and engineering faculty”, *Issues in Science and Technology Librarianship*, No. 64, available at: http://dx.doi.org/10.5062/F4H12ZXJ (accessed 7 June 2013).

Wagner, A.B. (2010), “Open access citation advantage: an annotated bibliography”, *Issues in Science and Technology Librarianship*, No. 60, available at: http://dx.doi.org/10.5062/F4Q81B0W (accessed 12 June 2013).

Wainer, J., Przibisczki de Olveira, H., and Anido, R. (2011), “Patterns of bibliographic references in the ACM published papers”, *Information Processing & Management*, Vol. 47 No. 1, pp. 135-42.

Wren, J.D. (2005), “Open access and openly accessible: a study of scientific publications shared via the internet”, *BMJ*, Vol. 330 No. 7500, p. 1128, available at: http://dx.doi.org/10.1136/bmj.38422.611736.E0 (accessed 28 June 2013).

Zhuang, Z., Wagle, R., and Giles, C.L. (2005), “What's there and what's not?: Focused crawling for missing documents in digital libraries,” in *JCDL’05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, June 7–11, 2005, Denver, Colorado, USA*, ACM: New York, pp. 301-10.

Appendix. Data Collection Specifications

Analyzed as Content Articles

Periodical articles, Conference papers, Editorials, Letters to editor, Prefaces, Introductions, Comment on an article, Awards, Memorials, Book reviews, 15 & 20 Years Ago Today, SIGGRAPH images, New products

Excluded from Analysis as Non-Content Results

Cover, Cover art, Editorial board / Committees / Organizing committee, About the authors, Guest editors lists, Reviewers, Title of the periodical or proceeding (even if Authors/Editors listed), Title page, Table of contents, Copyright notice, Call for papers, Call for nominations, Calendar, Conference announcement, Instructions for contributors, Indices: Author index, Article index, Subject index, Recent (Special) issues pages

Snapshot of Data Collected from the ACM Guide and the Google Search Results

INSERT Figure 4 (FORMAT OF DATA COLLECTION SPREADSHEET) HERE

Table I. Sample size per year

Year	Number in ACM Guide	Initial sample	Final sample	From periodicals	From conferences
2003	43568	130	119	54	65
2004	46955	140	131	57	74
2005	65635	196	175	73	102
2006	75844	226	214	70	144
2007	96138	287	265	85	180
2008	105104	314	291	118	173
2009	150654	450	427	108	319
2010	122055	364	345	113	232
Total	705953	2107	1967	678	1289

Figure 1. Google Scholar search result with shortcut to free full text circled in red and “All [number] versions” circled in orange for emphasis

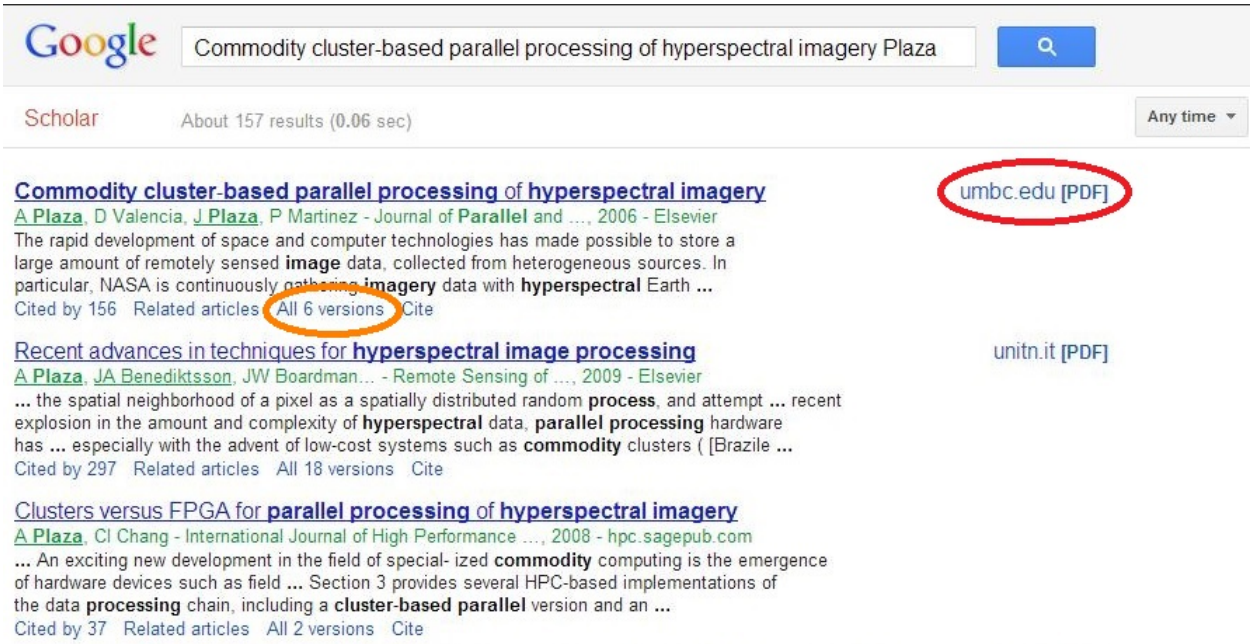

Figure 2. Google Scholar All [number] versions screenshot

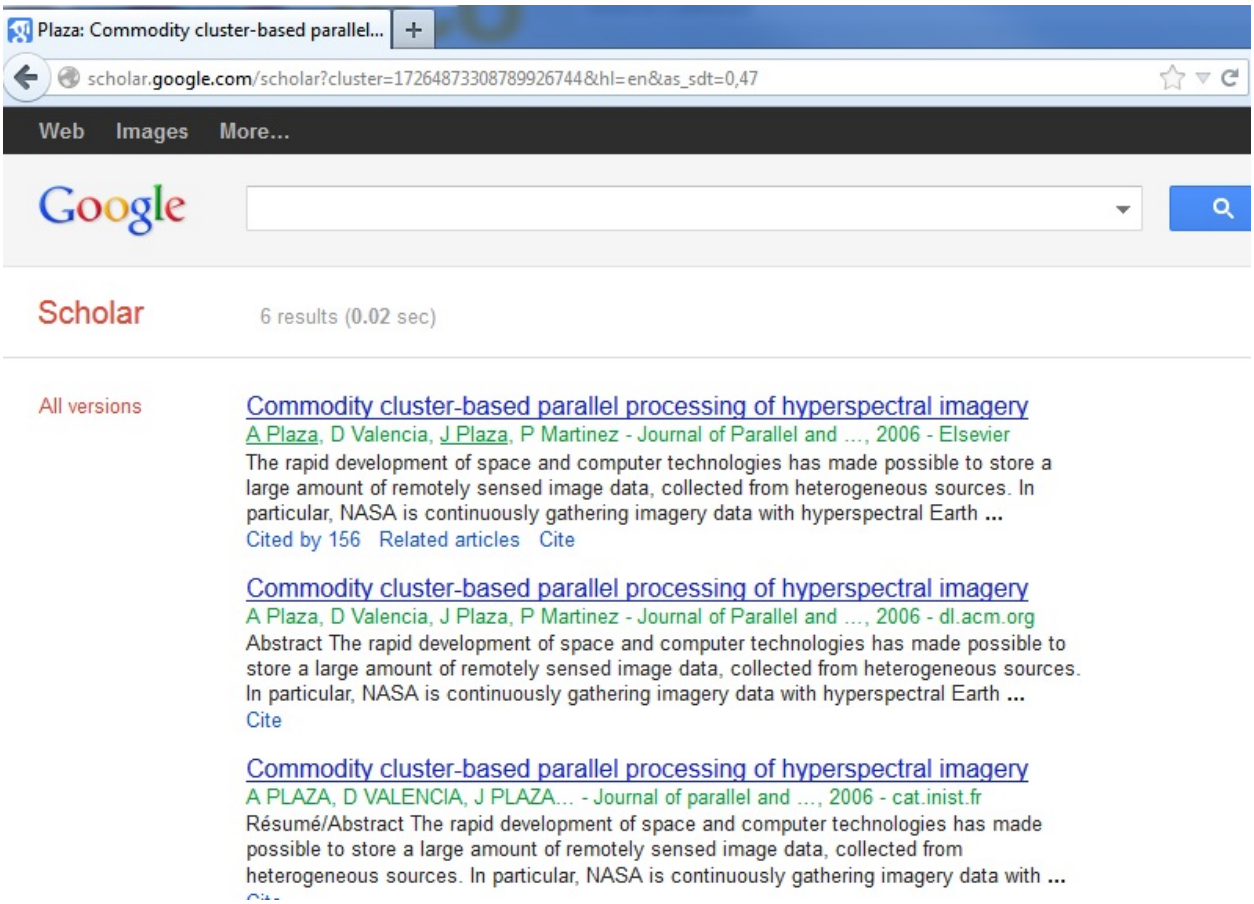

Figure 3. Percentage free full text by year

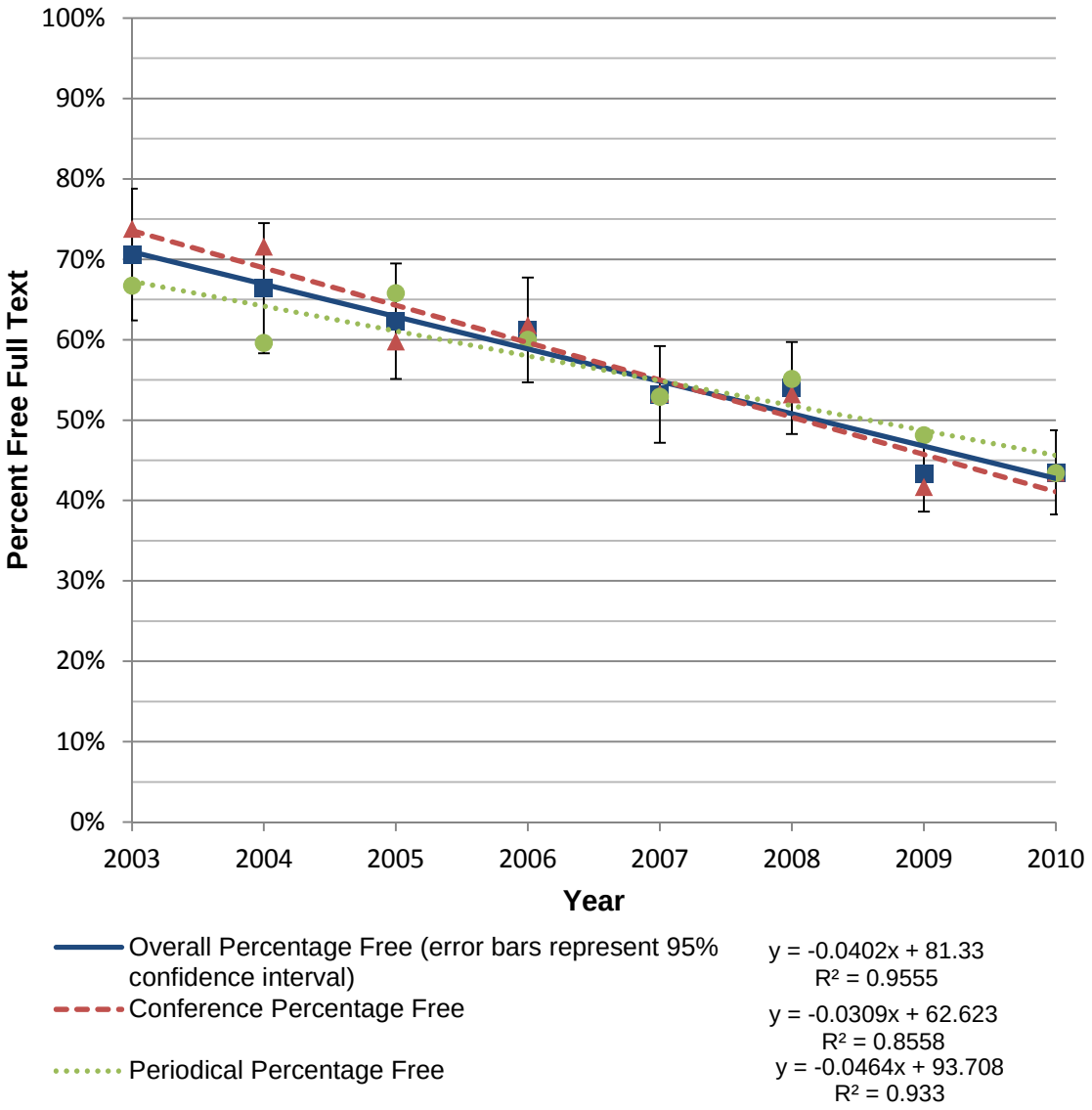

Figure 4. Format of data collection spreadsheet

Year	Random number	Guide bib result	Proceeding	Periodical	More than one search	Shortcut full text	Version full text	URL	No free copy	Not found at all	Exclude non-content
2006	64947	Editorial Board March 2006 Journal of Systems and Software , Volume 79 Issue 3		1							1
2006	65245	Comparison of in-network versus Staggered Multicast video distribution models H. Q. Guo, L. H. Ngoh, W. C. Wong, J. G. Tan March 2006 Multimedia Tools and Applications , Volume 28 Issue 3		1		1		http://www1.i2r.a-star.edu.sg/~guohq/PDF			
2006	65303	Extremely-Wide-Range Supply-Independent CMOS Voltage References for Telemetry-Powering Applications Amir M. Sodagar, Khalil Najafi March 2006 Analog Integrated Circuits and Signal Processing , Volume 46 Issue 3		1		1		http://deepblue.lib.umich			
2006	65649	Efficient PageRank approximation via graph aggregation A. Z. Broder, R. Lempel, F. Maghoul, J. Pedersen March 2006 Information Retrieval , Volume 9 Issue 2		1		1		http://delab.csd.auth.gr/~			
2006	66094	Software Construction, Part 2 Steve McConnell March 2006 IEEE Software , Volume 23 Issue 2		1	1				1	1	
2006	66259	Commodity cluster-based parallel processing of hyperspectral imagery Antonio Plaza, David Valencia, Javier Plaza, Pablo Martinez March 2006 Journal of Parallel and Distributed Computing , Volume 66 Issue 3		1		1		http://www.umbc.edu/rssi			
2006	66261	Quality and relevance of domain-specific search: A case study in mental health Thanh Tin Tang, Nick Craswell, David Hawking, Kathy Griffiths, Helen Christensen March 2006 Information Retrieval , Volume 9 Issue 2		1			1	http://citeseerx.ist.psu.ed doi=10.1.1.61.7172&rep=			
2006	67224	Adaptive beamforming and particle filtering Seth Benton, Andreas Spanias, Kai Tu, Harvey Thornburg, Gang Qian, Thanassis Rikakis February 2006 SPPRA'06: Proceedings of the 24th IASTED international conference on Signal processing, pattern recognition, and applications	1						1		
2006	67427	Multiresolution estimates of classification complexity and multiple subspace classifiers for understanding and solving complex recognition tasks Sameer Singh, Varun Kumar, Maneesha Singh February 2006 SPPRA'06: Proceedings of the 24th IASTED international conference on Signal processing, pattern recognition, and applications	1						1		