Information Processing and Management 48 (2012) 1116–1123

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Characterizing the frequency of repeated citations: The effects of journal, subject area, and self-citation

W.B. Lievers *, A.K. Pilkey

*Department of Mechanical and Materials Engineering, Queen's University, Kingston, Ontario, Canada K7L 3N6*

ARTICLE INFO

*Article history:*
Received 23 November 2010
Received in revised form 23 November 2011
Accepted 3 January 2012
Available online 3 March 2012

*Keywords:*
Repeated citation
Multiple mentions
Document search
Self-citation

ABSTRACT

Previous studies have repeatedly demonstrated that the relevance of a citing document is related to the number of times with which the source document is cited. Despite the ease with which electronic documents would permit the incorporation of this information into citation-based document search and retrieval systems, the possibilities of repeated citations remain untapped. Part of this under-utilization may be due to the fact that very little is known regarding the pattern of repeated citations in scholarly literature or how this pattern may vary as a function of journal, academic discipline or self-citation. The current research addresses these unanswered questions in order to facilitate the future incorporation of repeated citation information into document search and retrieval systems. Using data mining of electronic texts, the citation characteristics of nine different journals, covering the three different academic fields (economics, computing, and medicine & biology), were characterized. It was found that the frequency ($f$) with which a reference is cited $N$ or more times within a document is consistent across the sampled journals and academic fields. Self-citation causes an increase in frequency, and this effect becomes more pronounced for large $N$. The objectivity, automatability, and insensitivity of repeated citations to journal and discipline, present powerful opportunities for improving citation-based document search.



## 1. Introduction

Citation-based document searches are an extremely powerful tool that have been shown to provide results which complement other document retrieval methods, such as keyword searches (Kuper, Nicholson, & Hemingway, 2006). Document identification can be performed in either a reverse-chronological or forward-chronological fashion, depending on whether one works backwards from the references listed in some starting document, or whether one identifies subsequent works which cite that document.

One advantage of the reverse-chronological citation search is that it is generally closed-ended and well defined. Any given document cites only a finite number of earlier ones and each is contained in the reference list. Furthermore, the way in which each work is cited within the starting document can provide textual clues as to the reason it was cited and the importance ascribed to it by the author.

Conversely, forward-chronological citation searches are inherently open-ended. New works may be published on a daily basis which cite the original document. Identifying these works requires the use of secondary tools such as Thomson Reuters' *Web of Science*, *Scopus*, *Google Scholar*, or the publisher's website; however, these indexes of citing documents will rarely be

* Corresponding author. Present address: Center for Applied Biomechanics, University of Virginia, 4040 Lewis and Clark Drive, Charlottesville, VA 22911, United States. Tel.: +1 434 296 7288; fax: +1 434 296 3453.
*E-mail address:* lievers@virginia.edu (W.B. Lievers).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2012.01.009

complete. More significantly, if the document in question is cited frequently, the indexes may list hundreds or more citing documents, resulting in a situation well described by Voos and Dagaev (1976, p. 21) over thirty-years ago:

> Quite often an initial search of a subject provides more references than either the searcher has time to read, or the library has in its collection. In the latter case, since one does not really know the "relevance" of the items cited one either goes to an abstract journal hoping to find clarification, or orders it on interlibrary loan.

Although electronic access has simplified the process of locating and obtaining documents, the researcher must still read the paper in order to understand the nature of the link between the citing and the cited document.

Not all references are created equal, however. The challenge for document search and retrieval systems is in finding objective and automated ways to assess citation relevance. Research has shown that the section in which the citation is located (Voos & Dagaev, 1976), the text surrounding the citation (Bonzi, 1982), and the number of repeated citations within the citing document (Bonzi, 1982; Herlach, 1978) are all significant indicators of the relatedness between the citing and cited documents. This third method is the easiest to automate, particularly with the increased prevalence of electronic documents. For example, hyperlinks within HyperText Markup Language (HTML) documents can be mined to quantify repeated citations made within the text to each reference. This information could then be used to prioritize a citation search or could be combined with other metrics to develop improved document retrieval systems (Tang & Safer, 2008).

Previous research has established a strong link between repeated citations and relevance (Bonzi, 1982; Herlach, 1978; Tang & Safer, 2008); however, little is known about the patterns or characteristics of repeated citations in the scientific literature. This phenomenon must be better understood if it is to be incorporated into the next generation of search and retrieval systems. For example, early work in this area by Herlach (1978) simply divided references into two groups: those cited only once, and those cited two or more times. Yet works may be cited five, ten or even twenty times in the same document. There is little reason to believe that works cited twenty times will occur as frequently as those cited twice, nor is it expected that their relevance will be the same. A proper study relating relevance and repeated citation must be performed using a sample of documents which is representative of the underlying distribution; however, the nature of repeated citations in the scientific literature remains neither fully examined nor operationalized to add value to document retrieval systems.

The universality of the pattern of repeated citations must also be evaluated under various scenarios. It is well known that many scientometric measures – journal impact factors and Hirsch's h-index, for example – vary greatly within academic disciplines (Batista, Campiteli, Kinouchi, & Martinez, 2006; Ramírez, García, & Del Río, 2000) and can also be influenced by self-citations (Fassoulaki, Paraskeva, Papilas, & Karabinis, 2000; Schreiber, 2007). If the pattern of repeated citations is shown to be insensitive to these factors, it will facilitate the coupling of this objective measure with more subjective measures of document relevance.

The goal of the present work is to address these unanswered questions regarding the pattern of repeated citations in scholarly publications. Three journals were selected from each of three different topic areas (economics, computing, and medicine & biology). All the research articles published in 2008 for these nine journals were downloaded electronically in HTML format and custom software was employed to analyze the intra-document citation characteristics. Specifically, this work quantifies: (1) the frequency, $f$, with which references are cited $N$ or more times in the same document; (2) the effect of academic discipline on the $f$–$N$ relationship; and, (3) the effect of self-citation on the $f$–$N$ relationship. Improving our understanding of the nature of repeated citations within the scientific literature will ease the incorporation of this metric into improved document search and retrieval systems.

## 2. Methods

### 2.1. Document selection, retrieval and processing

Three research areas were selected in order to investigate the effect of discipline on repeated citations: economics, computing, and medicine & biology. For each research area, three journals were selected semi-randomly. The original pool of journals was limited to those that were published in English in 2008, those to which our institution (Queen's University) had electronic subscriptions, and those which made full article text available as HTML. To ease data mining, our selections were further limited to a single publisher (Elsevier) because of the consistent structuring of their HTML files. A minimum of 200 eligible articles in 2008 was also necessary. The nine journals selected are listed in Table 1.

Data mining was performed by downloading all articles published in 2008 for each journal in HTML format. Three criteria were used to determine article eligibility: having a minimum of three pages, containing at least one reference, and deemed to be a "research" article. This third criterion is difficult to summarize because of the different ways individual journals label and classify articles. Original research and topical reviews were included, whereas correspondence, commentaries, editorials, and introductions to special issues were excluded.

A custom program, written in the Python programming language, was employed to convert each HTML document to an Extensible Markup Language (XML) file. This file contained a hierarchical representation of the document structure: abstract, sections and subsections, list of references, footnotes, appendices, tables and figures. Citations made within each of these

**Table 1**
The three disciplines and nine journals selected for study.

| Discipline | Journal | Abbreviation |
|---|---|---|
| Economics | *Ecological Economics* | EE |
| | *International Journal of Production Economics* | IJoPE |
| | *Journal of Economic Behavior & Organization* | JoEBaO |
| Computing | *Computers & Mathematics with Applications* | CaMWA |
| | *Neurocomputing* | NC |
| | *Pattern Recognition* | PR |
| Medicine & biology | *Journal of Biomechanics* | JoB |
| | *Pain* | Pain |
| | *Virus Research* | VR |

components were identified using the hyperlinks within the HTML document. Once the structural XML file had been created, subsequent analysis of each document was easily performed using a second Python script.

### 2.2. Repeated citations by journal and research area

Each document was analyzed to count all citations made to each work listed in the references section. It should be noted that this method will only count explicit links within each document; implicit and non-hyperlinked explicit mentions will not be captured.

All the references in the eligible document within a journal were summed to generate counts, $c_i$, of the number of references which were cited $i$ times. From these data, the frequency ($f$) which any reference might be cited $N$ or more times within a single document was calculated as:

$$f(N) = \sum_{i=N}^{N_{\max}} c_i / \sum_{j=1}^{N_{\max}} c_j, \tag{1}$$

where $N_{\max}$ is the most times any reference is cited within any article.

The formulation of Eq. (1) was chosen for three reasons. First, it ensures that $f(1) = 1.0$ for all cases. Second, it guarantees that $f(N)$ will be continuously decreasing, particularly at large values of $N$. Finally, it eases truncation of a region at arbitrary values of $N$. This last factor facilitates comparisons with existing works which typically compare references cited once with those cited two or more times.

Since the number of articles and references varied dramatically by journal, the discipline behaviors were obtained by averaging the three $f(N)$ curves so that each journal was weighted equally. The total behavior of the nine journals was obtained similarly.

### 2.3. Effects of self-citation

The consistent formatting of references in the HTML files, resulting from the selection of journals produced by a single publisher, simplified the process of identifying self-citations. All reference authors appear in an ''initials surname'' format (e.g. ''D.E. Knuth'' and either ''Y.-P. Yu'' or ''Y.P. Yu''). The author names for each document were converted to this format, and a match with any references was labeled a self-citation. This subset of references was analyzed to see if authors are more likely to cite themselves repeatedly within a document.

Admittedly, this is a simplified approach. It is possible that some authors will cite the works of others who happen to share their initials and surname. In such cases self-citations will be overestimated. Conversely, inconsistent use of initials, changes in name, or transliteration issues will all lead to an underestimation. Since underestimation is more likely, these results will represent a convenient lower-bound to what might be expected given a more thorough method of identifying self-citations.

## 3. Results

### 3.1. Citation probability

The automated analysis method allowed for over 3000 articles, containing over 100,000 references, to be examined. Table 2 summarizes the number of articles and references per journal, as well as the average number of references per article, for the nine journals studied.

The most frequently cited reference or references were identified for each article, and the value $N_{\max}$ is equivalent to the total number of citations made to that reference. The frequency with which values of $N_{\max}$ occur are plotted in Fig. 1. Most

**Table 2**
Articles and reference statistics for each journal.

| Journal | Number of articles | Number of references | References/article | |
|---|---|---|---|---|
| | | | Mean ± st. dev. | Range |
| EE | 290 | 12,891 | 44.5 ± 24.1 | 8–192 |
| IJoPE | 356 | 12,333 | 34.6 ± 23.0 | 1–191 |
| JoEBaO | 216 | 7159 | 33.1 ± 18.2 | 3–136 |
| CaMWA | 578 | 11,817 | 20.5 ± 10.5 | 1–73 |
| NC | 379 | 10,824 | 28.6 ± 13.5 | 3–111 |
| PR | 306 | 9945 | 32.5 ± 14.3 | 10–116 |
| JoB | 471 | 13,960 | 29.6 ± 12.5 | 6–97 |
| Pain | 304 | 14,013 | 46.1 ± 16.2 | 14–106 |
| VR | 250 | 11,619 | 46.5 ± 35.7 | 13–334 |
| Total | 3150 | 104,561 | 33.2 ± 20.6 | 1–334 |

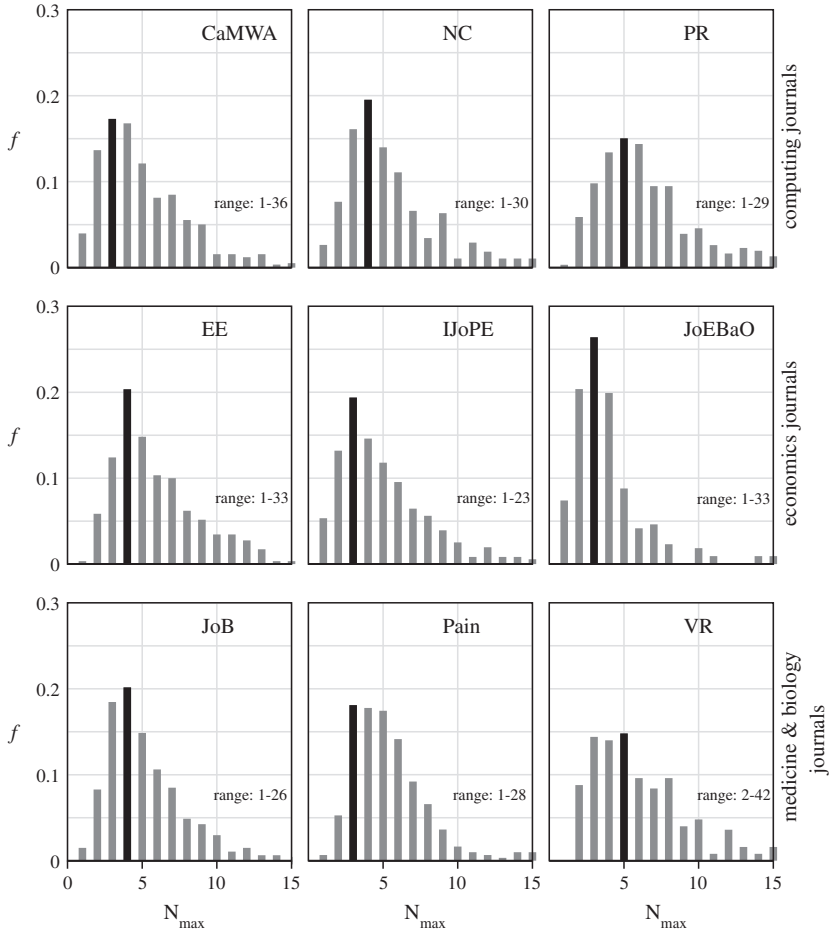


**Fig. 1.** Plots of frequency, *f*, with which the most frequency cited reference in an article is cited $N_{max}$ times. The black bar represents the most frequent $N_{max}$. The range of $N_{max}$ values is also shown.

articles typically have at least one reference which is cited 3–5 times. What is remarkable is that some references are cited as many 42 times in a single article, and every journal produced at least one article containing a reference cited over 20 times.

The frequency, *f*, of a particular work being cited within an article *N* or more times was calculated for each journal. These curves are presented in Fig. 2. The averaged curves are also shown for three disciplines (Fig. 3) and for the combined set of nine journals (Fig. 4).

The data was originally fit using a form of Lotka's law (Lotka, 1926), given as:

$$f = N^{-a}. \tag{2}$$

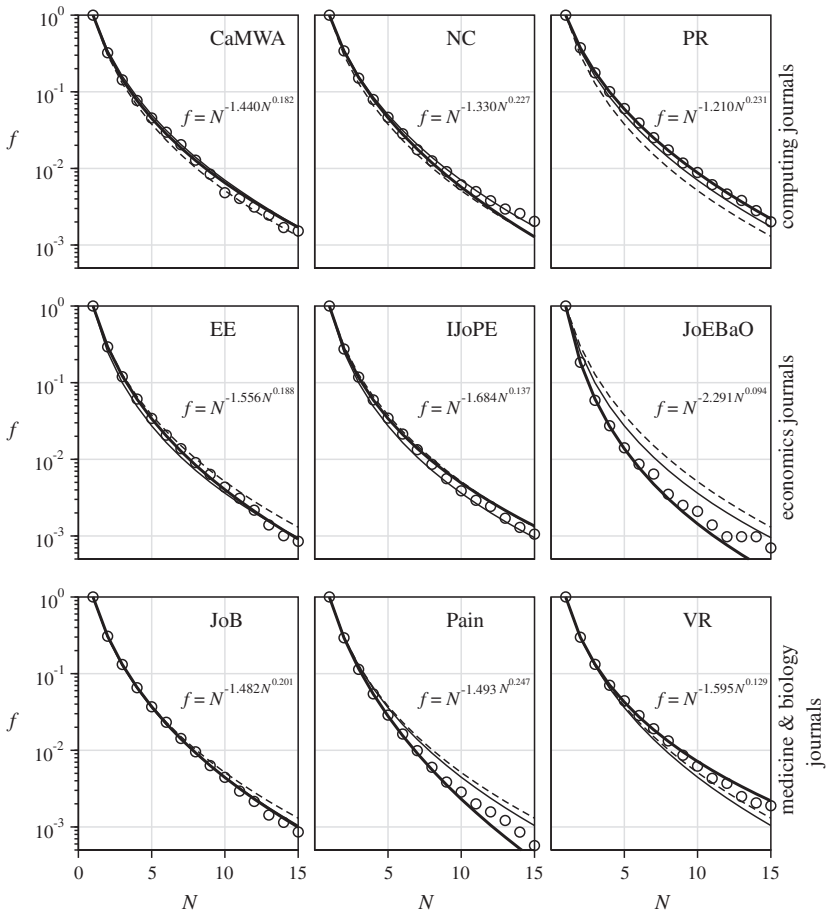


**Fig. 2.** Plots of frequency, *f*, with which references are cited *N* or more times. The dark solid line is the fit with Eq. (3), the solid line is the fit performed on the journal group, and the dashed line is the fit for the average of all nine journals.

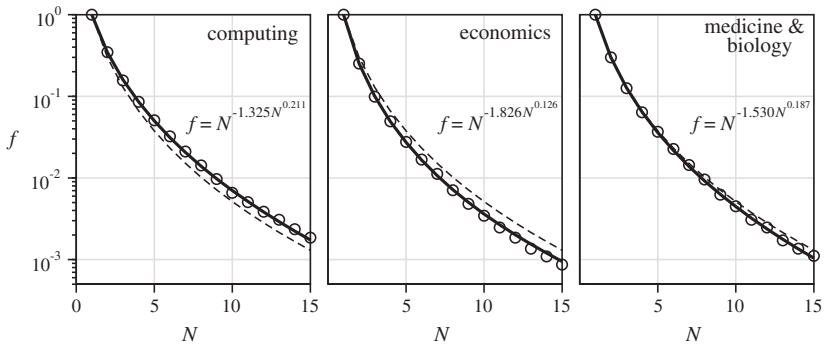


**Fig. 3.** Plots of frequency, *f*, with which references are cited *N* or more times for each of the three journal groups. The dark solid line is the fit for each journal group using Eq. (3). The dashed line is the fit for the average of all nine journals.

This was found to overestimate the probabilities for $N \geqslant 5$ (see Fig. 4). Instead, the data was found to be better described by an equation of the form:

$$f = N^{-a \cdot N^b}. \tag{3}$$

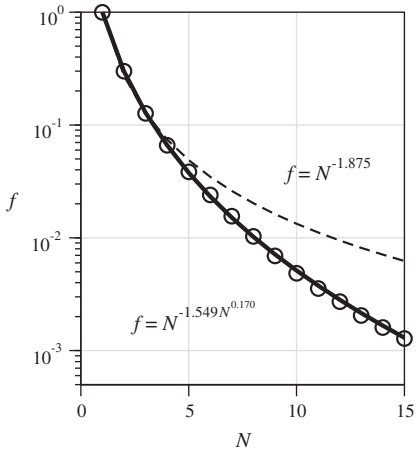


**Fig. 4.** Plots of frequency, $f$, with which references are cited $N$ or more times for all nine journals. The dark solid line is the fit using Eq. (3). The dashed line demonstrates the much poorer fit for large $N$ when using the Lotka's law of Eq. (2)

### 3.2. Self-citations

The subset of self-citations was identified and compared with all citations. Table 3 summarizes the number of articles which contained self-references and the total number of self-citations made for each journal. The percentages shown are relative to the total number of articles and references, as summarized in Table 2.

**Table 3**
Number of articles containing self-citations and the number of self-directed references (percentage of total).

| Journal | Number of articles | Number of references |
|---|---|---|
| EE | 210 (72.4%) | 925 (7.2%) |
| IJoPE | 222 (62.4%) | 587 (4.8%) |
| JoEBaO | 138 (63.9%) | 394 (5.5%) |
| CaMWA | 399 (69.0%) | 1589 (13.4%) |
| NC | 291 (76.8%) | 1163 (10.7%) |
| PR | 229 (74.8%) | 826 (8.3%) |
| JoB | 394 (83.7%) | 1764 (12.6%) |
| Pain | 289 (95.1%) | 1725 (12.3%) |
| VR | 227 (90.8%) | 1424 (12.3%) |
| Total | 2399 (76.2%) | 10,397 (9.9%) |

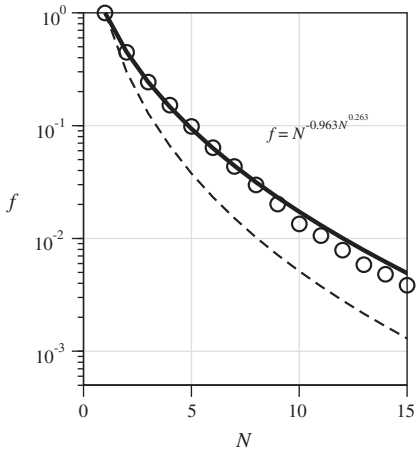


**Fig. 5.** Plots of frequency, $f$, with which self-cited articles are cited $N$ or more times for all nine journals. The dashed line is the average of all citations in all journals as shown in Fig. 4.

The *f*–*N* graph for the self-citations is shown in Fig. 5 and compared with the total behavior of all citations. The data indicate that self-references are more likely to be repeatedly cited; the effect becomes more pronounced for larger values of *N*.

## 4. Discussion

Forward-chronological reference searches are a powerful document retrieval tool, but the technique can result in an overwhelmingly large pool of potential articles. The repeated citation of references within an article has the potential to help prioritize the subsequent document examination by identifying those items which are truly relevant. Since some references might be cited as little as once, or as many as 30 or 40 times in an individual article, the incorporation of this information into a document search or retrieval system would greatly aid researchers.

The goal of the current work has been to evaluate the nature of repeated citations in scientific publications. It has been shown that the frequency, *f*, with which a reference will be cited *N* or more times, can best be expressed using Eq. (3). Only 3.8% of the references in the nine journals studied are cited five or more times ($N \geqslant 5$), and only 0.48% are cited 10 or more times ($N \geqslant 10$). Some references are even cited as many as 20 or more times, although such occurrences are exceedingly rare (0.05%).

Implicit in this work has been the assumption that a correlation exists between document relevance and repeated citation. While such a relationship has been repeatedly established by other researchers (Bonzi, 1982; Herlach, 1978; Tang & Safer, 2008), an explicit mathematical relationship has not been established herein. Nevertheless, the current study of repeated citations across three topic areas, nine journals, over 3000 articles, and over 100,000 citations, provides important information about the pattern of this phenomenon in a portion of the scientific literature. Knowledge regarding the statistical distribution is critical for establishing a smaller, representative set of documents from which an *N*-relevance relationship can be established. Future work is needed in this area.

Unfortunately, the task of "quantifying" relevance is inherently subjective and extremely laborious. Even once an appropriate *N*-relevance relationship has been established, it may not provide the sensitivity or specificity necessary for fully automated document search and retrieval. A simpler alternative, which could be implemented in a citation index, would be to provide users the option to sort their citation search results by the number of repeated citations. This could be implemented in the same manner by which results can currently be sorted by author name, year of publication, or other variables.

The current findings are consistent with the values reported by Herlach (1978). Her study indicated that 31.6% of references were cited two or more times ($N \geqslant 2$) within the same document, as compared with the value of 29.9% reported here. Furthermore, since her work was published in 1978, the near-equality of these two values suggests that the pattern of repeated citations has remained consistent for the past three decades. The present results also suggest that the pattern of repeated citations is relatively constant across the journals and research disciplines studied herein. If this stability could be demonstrated to exist more broadly in the research literature, then repeated citation analysis would have the tremendous advantage of not requiring the journal- or discipline-specific normalization that is common to other metrics (Batista et al., 2006; Ramírez et al., 2000).

The *f*–*N* relationship was found to be affected by self-citation in the sample documents. Although the practice of self-citation can be perceived as self-serving, there are a number of valid reasons for citing one's own research, particularly in fields that progress much more incrementally. Indeed, this is why self-citation has been found to be more prevalent in the physical sciences relative to the social sciences and humanities (Snyder & Bonzi, 1998). The present work is consistent with these findings (Table 3), as self-citation was found to be much less common in the economics journals (5.9% of references), compared with those of computing (11.0%) and medicine & biology (12.4%). Therefore, the repeated citation of self-references may be reflective of true relevance to the citing document as found by Tang and Safer (2008).

The current analysis included only research and review articles from three different journals in each of three different fields (computing, economics, and medicine & biology). While the pattern of repeated citations was found to be constant across these nine journals, the limited nature of the study sample must be recognized. Larger variability in the *f*–*N* relationship might be expected in other journals and other disciplines. The humanities and social sciences, for example, have different citation practices than the sciences studied herein. These potential differences require further examination. The *f*–*N* relationship is also be expected to differ in other publication types. For example, Budd (1986) has argued that the pattern of repeated citations will be different in books compared with journal articles. Dissertations and theses might also have different characteristics. This would need to be examined; however, since most documents in citation databases are journal articles and conference papers, publication-type differences may not be a significant issue for document retrieval purposes.

In the current analysis, each reference, and the frequency with which it was cited within a document, was treated independently of the other references within that document. This approach fails to account for author- and document-specific citing patterns. Tang and Safer (2008) reported that frequently-cited references were rated as highly relevant, particularly when the other references were infrequently cited. Normalization of the citation counts, perhaps by the average number of citations per reference, warrants further investigation.

It should also be noted that the focus of the current study was on the frequency with which references are cited repeatedly within journal articles. That is, a reverse-chronological search was employed to assess the behavior across a number of different documents. The *f*–*N* relationship reported herein will not be representative of the repeated citation characteristics of a forward-chronological search. Some documents may only ever be mono-cited, while others may have a large number of

documents which cite it repeatedly. This behavior may also be expected to change with time (Hooten, 1991). For example, repeated citation of a document may be more common shortly after publication and then become less frequent as the work is incorporated into the accepted knowledge of a discipline. These possibilities need to be examined in future work.

While the current work was performed with a view to document search and retrieval, repeated citations may also have applications in the quantification of scientific research. Citation information is currently used to generate journal impact factors (Garfield, 2006) and author indices such as the h-index (Hirsch, 2005). Since these techniques are insensitive to the nature of the citing relationship, information regarding repeated citations could complement and expand these metrics. A relevance score or weighting factor could be assigned to each reference, for example, based on the number of times, $N$, it is cited in a document. A modified h-index could then be proposed which incorporates these weightings into its calculation.

## 5. Conclusions

The present work has examined the pattern of repeated citations in nine academic journals taken from three research areas: economics; computing; and medicine & biology. The frequency, $f$, with which a reference is cited $N$ or more times was quantified and found to be consistent across the sampled journals and topic areas, but affected by self-citations.

Further research is needed to verify that these relationships extend beyond the limited subset of journal and disciplines studied herein, particularly those in the humanities and social sciences. A technique is also needed for generating a measure of the relatedness between the cited and citing documents. These developments would facilitate the inclusion of citation frequency information into bibliographic database systems, thereby providing a valuable tool for helping researchers prioritize citation-based searches.

## References


Batista, P. D., Campiteli, M. G., Kinouchi, O., & Martinez, A. S. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics, 68*(1), 179–189.

Bonzi, S. (1982). Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science, 33*(4), 208–216.

Budd, J. (1986). A citation study of American literature: Implications for collection management. *Collection Management, 8*(2), 49–62.

Fassoulaki, A., Paraskeva, A., Papilas, K., & Karabinis, G. (2000). Self-citations in six anaesthesia journals and their significance in determining the impact factor. *British Journal of Anaesthesia, 84*(2), 266–269.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association, 295*(1), 90–93.

Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? multiple mention of a reference as a characteristic of link between cited and citing article. *Journal of the American Society for Information Science, 29*(6), 308–310.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572.

Hooten, P. A. (1991). Frequency and functional use of cited documents in information science. *Journal of the American Society for Information Science, 42*(6), 397–404.

Kuper, H., Nicholson, A., & Hemingway, H. (2006). Searching for observational studies: what does citation tracking add to PubMed? A case study in depression and coronary heart disease. *BMC Medical Research Methodology, 6*, 4.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317–323.

Ramírez, A. M., García, E. O., & Del Río, J. A. (2000). Renormalized impact factor. *Scientometrics, 47*(1), 3–9.

Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *EPL, 78*(3), 30002.

Snyder, H., & Bonzi, S. (1998). Patterns of self-citation across disciplines (1980–1989). *Journal of Information Science, 24*(6), 431–435.

Tang, R., & Safer, M. A. (2008). Author-rated importance of cited references in biology and psychology publications. *Journal of Documentation, 64*(2), 246–272.

Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or, did we op. cit. your idem? *Journal of Academic Librarianship, 1*(6), 19–21.