The Distribution of References Across Texts: Some Implications for Citation Analysis

Ying Ding, Xiaozhong Liu, Chun Guo, Blaise Cronin
School of Library and Information Science, Indiana University, 1320 E. 10th St, Bloomington, IN. 47405, USA
{dingying, liu237, chunguo, bcronin}@indiana.edu

Abstract

In citation network analysis, complex behavior is reduced to a simple edge, namely, node A cites node B. The implicit assumption is that A is giving credit to, or acknowledging, B. It is also the case that the contributions of all citations are treated equally, even though some citations appear multiply in a text and others appear only once. In this study, we apply text-mining algorithms to a relatively large dataset (866 information science articles containing 32,496 bibliographic references) to demonstrate the differential contributions made by references. We 1) look at the placement of citations across the different sections of a journal article, and 2) identify highly cited works using two different counting methods (CountOne and CountX). We find that 1) the most highly cited works appear in the Introduction and Literature Review sections of citing papers, and 2) the citation rankings produced by CountOne and CountX differ. That is to say, counting the number of times a bibliographic reference is cited in a paper rather than treating all references the same no matter how many times they are invoked in the citing article reveals the differential contributions made by the cited works to the citing paper.

Introduction

An author's motivations for citing and the reasons why an author distributes citations across any given text in a particular fashion can be inferred, with greater or lesser confidence, but, ultimately, not known. At the risk of stating the obvious, citing is an inherently subjective act. Nevertheless, there exists a set of norms—Cronin (2004, p. 43) speaks of "the normative ghost in the machine"—and procedural standards to which scientists typically adhere (e.g., Cronin, 1984; Small, 1976). The process of selecting and dressing a work with references is far from random (Cronin, 1981; Small, 2011a). As Small (1978, p. 328) states, "Referencing … is a labeling process. The language pointed to by the footnote number labels or characterizes the document cited—or, in other words, constitutes the author's interpretation of the cited work. In citing a document an author is creating its meaning, and this … is a process of symbol making." Citing behavior may indeed be subjective in character, but large-scale citation analyses are widely used in research evaluation exercises (Garfield, 1972; Hirsch, 2005). Despite persistent concerns about validity and reliability (e.g., MacRoberts & MacRoberts, 2010) we align ourselves with the pragmatic White (1990, p. 91): "When one sees that scores, hundreds, and even thousands of citations have accrued to a work, an author, a set of authors, it is difficult to believe that all of them are suspect."

Citation analysis assumes that (citing) author A has been influenced by the work of (cited) author B, though without any attempt to specify the strength or direction of that influence. Additionally, it is assumed that each reference has made an equal contribution to the citing article. The pioneering work of Pinski and Narin (1976) inspired researchers to view citing behavior as networked influence. Researchers have started to apply PageRank to measure journal influence (Bollen, Rodriguez, & Van de Sompel, 2006; Bergstrom, 2007) and author influence (Ding, 2011). In short, citations should not be weighted equally: edges coming from, say, highly cited authors, journals or papers should, arguably, be granted greater weight.

In citation network analysis, citing behavior is construed as an edge in which node A cites node B—nodes can be authors, papers, journals or institutions (Pinski & Narin, 1976). It is sometimes assumed that A's citing of B is equivalent to A's voting for/giving credit to/endorsing B. In traditional bibliometric analysis, the contribution of each citation is weighted equally. With advances in computing capability and the growth of digital libraries and repositories, it is possible to develop scalable text-mining algorithms to extract associations hidden in large document collections. These developments herald the next generation of citation analysis: content-based citation analysis.

Content-based citation analysis can be conducted at different levels: 1) at the syntactic level (the location of references), where citations are differentiated based on the structural features of journal articles; 2) at the semantic level (the meaning of references), where citations are analyzed based on the nature of the contributions they make (e.g., significant/trivial, positive/negative) to the argument or development of the citing article by using advanced text-mining and natural language processing technologies. Here, we explore content-based citation analysis at the syntactic level.

In this paper, the terms citation and reference are not used synonymously. Reference refers to the works mentioned in the reference section or bibliography of a journal article. A reference may be mentioned once or multiply in an article. Each mention is considered a citation, so, for example, a single reference could have three citations if mentioned three times in a citing paper. Thus, there are two ways of counting citations: 1) count each reference once no matter how many times it is mentioned in an article (we call this CountOne, which is the traditional bibliometric approach); and 2) consider each reference as having been cited X times if it is mentioned a total of X times in an article (we call this CountX).

Here we apply text-mining algorithms to a relatively large dataset (866 articles containing 32,496 unique references) to reveal the differential contributions made by references by: 1) looking at the distribution of citations across the different sections of a paper, and 2) comparing highly cited works using the aforementioned counting methods. The paper is organized as follows: the Related Work section surveys the state-of-the-art on content-based citation analysis; the Methods section describes the dataset and data processing techniques we used; the Results/Discussion section interprets the findings; and the Conclusion summarizes our results and identifies avenues for future research.

Related Work

Sociological grounding of citing behavior

In the late 1970s, sociologists and bibliometricians devoted considerable effort to examining the nitty-gritty of citation practice in order to justify (or contest) the use of citations in evaluative bibliometrics. They adopted a variety of approaches, such as examining the context (surrounding text) of citations (Chubin & Moitra, 1975), analyzing the function and quality of citations (Moravcsik & Murugesan, 1975), and developing classification schemes for citations (Lipetz, 1965). Small (1982) and Cronin (1984) have both provided overviews of citation classification schemas. Lipetz (1965) identified 29 different reasons for citing and grouped them into four clusters: 1) original scientific contribution or intent of citing paper; 2) contribution of citing paper other than original scientific contribution; 3) identity or continuity relationship of citing paper to cited paper; and 4) disposition of the scientific contribution of the cited paper to the citing paper. Frost (1979) develop a taxonomy appropriate to the humanities: 1) documentation of primary sources to support an opinion or factual statement; 2) documentation of secondary sources to approve or disapprove of cited scholars; and 3) documentation of sources, either primary or secondary, to refer to further reading or provide bibliographical information on a specific edition. Moravcsik and Murugesan (1975) categorized citations as follows: conceptual/ operational, evolutionary/juxtapositional, organic/perfunctory, and confirmative/negational. Chubin and Moitra (1975) established a tree hierarchy distinguishing between confirmative (four types) and negational (two types) citations. McCain and Turner (1989) conducted a bibliometric analysis of citation choice patterns within the field of Molecular Genetics by focusing on the aging patterns of individual journal articles. Chubin and Moitra (1975) identified six mutually exclusive categories of citation. Oppenheim and Renn (1978) and Spiegel-Rosing (1977) distinguished between methodological (e.g., providing data, developing methods, etc.) and general contributions (e.g., historical background of a subject domain) of cited works. With the exception of a few small-scale investigations (e.g., Cronin, 1981; Finney, 1979; Small, 2011b), most bibliometric studies have ignored the actual distribution of citations across the constituent parts (e.g., Introduction, Results) of a text.

Syntactic content-based citation analysis

The structural layout of research papers is fairly standardized (Suppe, 1998), though evolving in terms of both functionality and features (see: http://www.articleofthefuture.com/). It typically follows a sequence such as this: a) Abstract, b) Introduction, c) Theoretical Background, d) Experimental or Observational Techniques, e) Samples, f) Data Analysis, g) Results or Observations, h) Discussion, i) Summary/Conclusions, j) Acknowledgments, k) References, and l) Appendices. Frequently, b and c are combined, d-f can be viewed as the Methodology part, and f-h as the Results section. Sections about methods, data, interpretations, and the replicability of observational claims of a paper are most important when it comes to assessing whether the new knowledge claims can be integrated into the discipline's shared knowledge base (Suppe, 1998).

Voos and Dagaev (1976) analyzed the contributions of citations in relation to the component parts of articles (Introduction, Methodology, Discussion, etc.). They found that highly cited articles appeared most often in the Introduction and argued that the contribution of a cited reference can be calculated based on the number of times it is cited and the location of those re-citations in the citing article. In similar vein, Herlach (1978) maintains that if a work has been cited in the Introduction or Literature Review section and is mentioned again in the Methodology or Discussion sections, it is likely that it makes a greater overall contribution to the citing paper than others that have been mentioned only once. Bonzi (1982) examined stylistic aspects of citation, distinguishing between a number of broad categories: those not specifically mentioned in the text (e.g., "several studies have dealt with …"); those barely mentioned (e.g., "Smith has studied the impact of …"); direct quotation or discussion point (e.g., "Smith found that …"); and two or more quotations or discussion points in the text.

Semantic content-based citation analysis

Methods used for semantic content-based citation analysis have been limited to interviews (e.g., Vinkler, 1987) and the manual processing of full-text articles (e.g., McCain & Turner, 1989; Spiegel-Rosing, 1977). Small (1978) was among the first to study citation context by viewing citations as symbols standing for concepts or methods, which echoes Garfield's (1977) semantic notion of cited documents as subject headings in an indexing system. Small (1978, p. 334) examined citation contexts (i.e., 2-3 sentences around the points where citations appeared in the text) for a set of highly cited articles in chemistry and found that most of the articles were not "research front papers" but "well-established instructions on how to carry out certain basic operations at the lab bench or at the desk". He found that many highly cited documents in chemistry have uniform or standardized usage and meaning.

Recently, Teufel, Siddharthan and Tidhar (2006) proposed a reliable citation function annotation schema so that a supervised machine-learning framework could automatically classify citation functions using both shallow and deep natural language features. They created four top-level categories (explicit statement of weakness, contrast or comparison with other work, agreement/usage/compatibility with other work, and a neutral category) and automatically labeled each citation with a single category. They tested their approach on 360 conference papers and found a strong relationship between citation function and sentiment classification. More recently, Small (2011) analyzed citation sentiments based on the text within the vicinity of references appearing in scientific papers. He defined citation context as one to three sentences around the citation. In this study of 81 full-text, co-cited papers in the organic thin-film transistors domain, he used an average of 1.6 sentences surrounding the reference.

This paper 1) analyzes full-text articles using automatic approaches to detect distinctive citation distributions in the different sections of citing papers, and 2) counts the total number of times a cited reference is invoked in the citing article.

Methods

Dataset

We used the *Journal of the American Society for Information Science and Technology* (*JASIST*) to generate our dataset, assembling a total of 866 full-text research articles for the period 2000 to 2011: document types such as "editorial" and "book reviews" were excluded.

Data Processing

The challenge was to identify a reference in a full-text article and associate each invocation of that reference with a specific section of the article, namely, Abstract, Introduction, Literature Review, Methodology, Result, and Conclusion and Future Work. The full-text articles we collected were in HTML format. We used HTML metadata (e.g., HTML tags, such as <a href="...." rel="references...">) to identify a reference in a full-text article. This was a straightforward process. However, it is a nontrivial task to figure out where a reference is cited since HTML tags for section names differ significantly. For example, some articles employed <h3>, i.e. , *<h3> Introduction </h3>*, to emphasize a section name, but some not. There is no consistent way to represent sections in a given full-text article using HTML tags. To address this problem, we used two approaches: automatic extraction and manual coding.

Automatic Extraction

A program written by regular expression was used to automatically capture the pattern of special HTML tags (e.g., <h1>, <h2>, <h3>, etc.) and extract the section names from full-text articles in HTML format. We also used rules to verify whether the section names extracted were correct. For instance, the character length of the section name should be no more than 300, while the word token length for each section should be no more than 15; this helps reduce noise. First, for each section, the word tokens were extracted from the section name and stop words (e.g., "and", "to", or "the") were removed from the section names. Next, a snowball-stemming algorithm was applied to obtain each word's morphological root form. For example, based on the stemming algorithm, "research", "researches", "researching" and "researcher" were projected onto the word stem "research". Table 1 lists the most frequently appearing stemmed word tokens in each section.

Table 1: The most frequently stemmed tokens in each physical section

Section	Section Name (most frequent stemmed tokens)
Abstract	abstract
Introduction	introduct
Literature Review	background, literatur, framework, previo, relat, measur, method
Methodology	research, studi, data, design, discuss, hypothes, data, approa
Result	discussion, results, experi, result, user, system, evalu, discu
Conclusion/Future Work	conclus, theori, summar, limit, evalu, discu, analy, find, futur

Based on Table 1, we manually generated rules to improve the accuracy of the automatic extraction. For example, for the Literature Review section, if the section name appears after the second section and before the fifth section of the HTML full-text article, and the section name has one of the following word tokens {liter, backgrou, relat, resear, review, previo, studi}, we consider it to be the Literature Review section. We ended up with 866 Abstract sections, 863 Introduction sections, 389 Literature Review sections, 1,081 Methodology sections, 902 Result sections, 680 Conclusion sections, and 2,085 unknown sections. In total 6,866 different sections were automatically extracted. The number of Methodology sections is greater than the number of publications in the test collection because some articles have more than one section that deals with methodological issues. For example, a paper could have separate sections about "research questions" and "experimental design". The number of Literature Review sections is less than the number of publications because some review articles do not have clear-cut Review sections. In some cases, the Literature Review is effectively embedded in either the Introduction or Methodology section. As a result, many sections are difficult to classify.

Manual Coding

Manual coding was conducted to cross-validate the automatically extracted sections and to verify the 2,085 unknown sections. Two information science doctoral students manually validated the 6,866 sections extracted from the 866 full-text articles. First, we trained the coders for three weeks in face-to-face meetings. During the training practice, the students coded 366 articles together but independently with an agreement rate of 81%. Second, after training, each one coded 250 articles independently. If the students could not decide which section the automatically extracted section belonged to, they marked it "unknown". If unsure, they could flag the case and ignore the paper. Ultimately, the coders verified the following: 847 Abstract sections, 831 Introduction sections, 373 Literature Review sections, 758 Methodology sections, 559 Result sections, and 821 Conclusion and Future Work sections. These sections finally verified by the coders were used in our study.

Results and Discussion

Overview

The standard approach counts each reference once no matter how many times it is mentioned in a text (CountOne). In this paper we consider each reference as having been cited X times if it is mentioned a total of X times (CountX). The *Journal of the American Society for Information Science and Technology* published a total of 866 research articles containing 32,496 references (CountOne) and 53,017 mentions (CountX) during the period 2000-2011. The highest mention is a paper titled with "Stability of INEX 2007 Evaluation Measures" which was a workshop paper published in the Proceedings of the 7th NTCIR Workshop on Evaluation of Information Access Technologies in 2008. This paper was mentioned 39 times by one paper. The most highly cited article/book is Salton and McGill's popular *Introduction to Modern Information Retrieval*, which was cited 51 times (CountOne) but mentioned 87 times (CountX). The distribution of references using CountOne and CountX follows a power-law distribution with few titles highly ranked and most found in the long-tail; for instance, 94% of references were cited less than three times (CountOne) and 78% of references were mentioned less than three times (CountX).

Citation Location

Table 2 displays the distribution of the highly cited articles using CountOne across the different article sections. Because the number of different sections in the 866 articles varies, for comparison purpose, the average citation frequency in a specific section was used, which equals the total number of references which have been cited X times in a specific section divided by the number of times this specific section was identified in the 866 articles. For example, for a reference cited twice, the average citation frequency in the Literature Review section equals 1.330 (496/373), and the average citation frequency for the Introduction section is 0.918 (763/831). So, more references cited twice are located in Literature Review section than in Introduction section.

Table 2 shows that highly cited papers (using CountOne, with citation frequency >=10) appear more in the Introduction section (0.017), followed by the Literature Review section (0.013), and the Methodology section (0.009). For less cited papers (using CountOne, with citation frequency=2), more appear in the Literature Review section (1.330), followed by the Methodology section (0.982), and the Introduction section (0.918). The Introduction, Literature Review, and Methodology sections account for most of the citations, while, the Abstract, Result, and Conclusion and Future Work plots follow a stable course.

Table 2: Reference frequency distribution across article sections (CountOne)

Citation Frequency (X)	X=2	X=3	X=4	X=5	X=6	X=7	X=8	X=9	X>=10
Abstract (847)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Introduction (831)	0.918	0.243	0.085	0.037	0.020	0.007	0.005	0.011	0.017
Literature Review (373)	1.330	0.351	0.075	0.046	0.029	0.019	0.003	0.005	0.013
Methodology (758)	0.982	0.212	0.087	0.033	0.018	0.021	0.005	0.009	0.009
Result (559)	0.268	0.039	0.016	0.002	0.007	0.000	0.004	0.000	0.000
Conclusion/Future Work (821)	0.312	0.049	0.021	0.005	0.001	0.000	0.001	0.000	0.001

Note: Cell value is the average citation frequency. Average citation frequency in a specific section = the total number of references which have been cited X times in a specific section/the total number of this specific section identified in the 866 articles

Table 3 displays the citation frequency distribution across the different article sections using CountX. For highly cited articles (using CountX, with citation frequency >=10), a majority is found in the Methodology section (0.113), followed by the Literature Review section (0.086) and the Introduction section (0.064). This is considerably different from Table 2. For the less cited articles (using CountX, with citation frequency=2), a majority is located in the Literature Review section (2.928), followed by the Introduction (1.714), and the Methodology section (1.653). In general, a majority of citations is located in the Literature Review, Methodology, and Introduction sections, while the Abstract, Result and Conclusion and Future Work sections have fewer.

Table 3: Reference frequency distribution across article sections (CountX)

Citation Frequency	X=2	X=3	X=4	X =5	X =6	X =7	X =8	X =9	X >=10
Abstract (847)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Introduction (831)	1.714	0.588	0.285	0.132	0.082	0.049	0.042	0.022	0.064
Literature Review (373)	2.928	1.190	0.496	0.292	0.180	0.070	0.062	0.038	0.086
Methodology (758)	1.653	0.600	0.280	0.169	0.102	0.061	0.040	0.028	0.113
Result (559)	0.590	0.208	0.079	0.041	0.020	0.011	0.009	0.004	0.020
Conclusion/ Future Work (821)	0.635	0.194	0.074	0.035	0.015	0.006	0.002	0.007	0.006

Note: Cell value is the average citation frequency. Average citation frequency in a specific section = the total number of references which have been cited X times in a specific section/the total number of this specific section identified in the 866 articles

Table 4 shows the placement of the top 10 most highly cited references across the different article sections. Broadly speaking, information retrieval, information seeking and information visualization are the three most highly cited topics in the *JASIST* data set. The Introduction tends to cite work that sets the scene and/or introduces research questions, and overlaps with the Literature Review section which transitions to the Methodology section. The Methodology section typically cites specific papers discussing particular methods. The Results section has elements in common with the

Introduction, the Literature Review and the Methodology sections. It is not uncommon for the Results section to connect the study's findings to the prior research cited in both the Introduction and Literature Review. The citation profile of the Conclusion and Future Work section is quite different from the other sections.

Table 4: Most highly cited references across different article sections

Section	Top 10 highly cited references (CountOne)
Introduction	1. Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*: McGraw-Hill, Inc. (Times cited: 15) 2. Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths. (Times cited: 12) 2. Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361-371. (Times cited: 12) 2. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine, *Proceedings of the Seventh World Wide Web Conference* (pp. 107-117). Brisbane, Australia. (Times cited: 12) 2. Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences, 16*, 317-323. (Times cited: 12) 6. Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America, 102*(46), 16569-16572. (Times cited: 11) 6. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407. (Times cited: 11) 6. Ingwersen, P. (1998). The calculation of web impact factors. *Journal of documentation, 54*(2), 236-243. (Times cited: 11) 6. Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407-424. (Times cited: 11) 10. Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3-50. (Times cited: 10) 10. Dunman, S. (1998). Seeking meaning: A process approach to library and information services. *Journal of the American Society for Information Science, 47*(3), 249-250. (Times cited: 10)
Literature Review	1. Marchionini, G. (1997). *Information seeking in electronic environments*: Cambridge University Press. (Times cited: 14) 2. Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361-371. (Times cited: 13) 3. Dunman, S. (1998). Seeking meaning: A process approach to library and information services. *Journal of the American Society for Information Science, 47*(3), 249-250. (Times cited: 11) 4. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604-632. doi: 10.1145/324133.324140 (Times cited: 10) 4. Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation, 56*(5), 540-562. (Times cited: 10) 6. Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths. (Times cited: 9) 6. Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern information retrieval*: Addison-Wesley Longman Publishing Co., Inc. (Times cited: 9) 8. Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45*(3), 149-159. (Times cited: 8) 9. Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6-12. doi: 10.1145/331403.331405 (Times cited: 7) 9. Jörgensen, C. (1998). Attributes of images in describing tasks. *Information Processing & Management*, *34*(2), 161-174. (Times cited: 7) 9. Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*: McGraw-Hill, Inc. (Times cited: 7) 9. Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), 602-615. (Times cited: 7) 9. Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology (ARIST), 29*, 3-48. (Times cited: 7) 9. Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*(9), 810-832. (Times cited: 7)
Methodology	1. Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*: McGraw-Hill, Inc. (Times

	cited: 30) 2. Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths. (Times cited: 24) 3. Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137. (Times cited: 14) 4. Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern information retrieval*: Addison-Wesley Longman Publishing Co., Inc. (Times cited: 13) 5. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523. (Times cited: 12) 6. Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill. (Times cited: 10) 7. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46. (Times cited: 9) 7. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47. (Times cited: 9) 9. Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3-50. (Times cited: 8) 9. Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications. (Times cited: 8) 9. Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *56*(1), 71-90. (Times cited: 8) 9. Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features, *Proceedings of the 10th European Conference on Machine Learning* (pp. 137-142). Chemnitz, Germany: Springer-Verlag. (Times cited: 8)
Results	1. Lawrence, S., & Giles, C. (1999). Accessibility of information on the web. *Nature*, *400*(6740), 107-9. (Times cited: 8) 2. White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, *49*(4), 327-355. (Times cited: 7) 3. Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6-12. doi: 10.1145/331403.331405 (Times cited: 6) 3. Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern information retrieval*: Addison-Wesley Longman Publishing Co., Inc. (Times cited: 6) 3. Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill. (Times cited: 6) 6. Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths. (Times cited: 5) 6. Hair, J., Black, B., Babin, B., Anderson, R., & Tatham, R. (2005). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice-Hall. (Times cited: 5) 8. Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information processing & management*, *39*(3), 445-463. (Times cited: 4) 8. Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *Computer*, *35*(3), 107-109. (Times cited: 4) 8. Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill. (Times cited: 4) 8. Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing & Management, 36*(2), 207-227. (Times cited: 4) 8. Spink, A., Wolfram, D., Jansen, M. B., & Saracevic, T. (2000). Searching the web: The public and their queries. *Journal of the American society for information science and technology*, *52*(3), 226-234. (Times cited: 4) 8. Lawrence, S., & Giles, C. L. (1998). Searching the world wide web. *Science*, *280*(5360), 98-100. (Times cited: 4) 8. Neuendorf, K.A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage. (Times cited: 4) 8. Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, *99*(8), 5207-5211. (Times cited: 4)
Conclusion and Future Work	1. Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361-371. (Times cited: 10) 2. Dunman, S. (1996). Seeking meaning: A process approach to library and information services. *Journal of the American Society for Information Science, 47*(3), 249-250. (Times cited: 7) 3. Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407-424. (Times cited: 6) 4. Hill, J. R. (1996). Information seeking in electronic environments. *Journal of the American Society for*

	Information Science, 47(5), 405-406. (Times cited: 5)
	4. Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), 602-615. (Times cited: 5)
	6. Koenemann, J., & Belkin, N. J. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground* (pp. 205-212). ACM. (Times cited: 4)
	6. Fidel, R., Davies, R. K., Douglass, M. H., Holder, J. K., Hopkins, C. J., Kushner, E. J., ... & Toney, C. D. (1999). A visit to the information mall: Web searching behavior of high school students. *Journal of the American Society for Information Science*, *50*(1), 24-37. (Times cited: 4)
	6. Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation, 56*(5), 540-562. (Times cited: 4)
	6. Moed, H. F. (2005). *Citation analysis in research evaluation*. New York: Springer. (Times cited: 4)
	6. Belkin, N. J. (1984). Cognitive models and information transfer. *Social Science Information Studies*, *4*(2), 111-129. (Times cited: 4)
	6. Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3-50. (Times cited: 4)
	6. Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American society for information science and technology*, *59*(6), 983-1001. (Times cited: 4)
	6. Ingwersen, P. (1992). *Information retrieval interaction* (Vol. 246). London: Taylor Graham. (Times cited: 4)
	6. Saracevic, T. (1999). Information science. *Journal of the American Society for Information Science, 50*(12), 1051-1063. (Times cited: 4)
	6. Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge university press. (Times cited: 4)
	6. Layne, S. S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science, 45*(8), 583-588. (Times cited: 4)
	6. Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology, 37*(1), 413-464. (Times cited: 4)
	6. Fidel, R. (1997). The image retrieval task: implications for the design and evaluation of image databases. *New Review of Hypermedia and Multimedia*, *3*(1), 181-199. (Times cited: 4)
	6. Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context*. Dordrecht, Netherlands: Springer. (Times cited: 4)
	6. Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45*(3), 149-159. (Times cited: 4)

CountOne vs. CountX

As mentioned in the Introduction, traditional citation analysis treats each bibliographic citation in the reference list of a paper equally, irrespective of the number of times each one is cited in the body of the text. We, like Voos and Dagaev (1976), maintain that a multiply cited reference probably plays a more significant role (has made a potentially greater contribution) than a reference that has been cited only once.

In *JASIST*, the 32,496 unique references generate 53,017 mentions, an average of roughly 1.6 citations per reference. Table 5 lists the top 10 most highly cited works in *JASIST* (2000-2011) using the two different counting methods introduced earlier (CountOne vs. CountX): 30% of them are different. We can see that top ranked articles based on CountX are not necessarily the top cited articles based on CountOne. For example, Silverstein and coauthors' paper, *Analysis of a very large web search engine query log*, was cited 16 times (CountOne), but was mentioned 50 times (CountX). This means that their paper was mentioned on average 2.63 times for each citation. Furthermore, we found that this paper was cited 5 times and mentioned 12 times in the Methodology section, and cited 6 times and mentioned 15 times in the Result section. This demonstrates the significant influence of this paper for methodology development and results comparison. Similarly, for Barry's paper, which was mentioned more frequently in the Literature Review section (cited 8 times, mentioned 19 times) and the Methodology section (cited 4 times and mentioned 12 times). It is significant that CountX managed to bring these two papers up to the top 10 list.

Table 5: Top 10 most highly cited works using CountOne and CountX

CountOne	CountX
Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*: McGraw-Hill, Inc. (cited 51 times)	Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern information retrieval*: Addison-Wesley Longman Publishing Co., Inc. (mentioned 93 times)
van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworths. (cited 51 times)	Dunman, S. (1996). Seeking meaning: A process approach to library and information services. *Journal of the American Society for Information Science, 47*(3), 249-250. (mentioned 90 times)
Dunman, S. (1996). Seeking meaning: A process approach to library and information services. *Journal of the American Society for Information Science, 47*(3), 249-250. (cited 32 times)	Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*: McGraw-Hill, Inc. (mentioned 87 times)
Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern information retrieval*: Addison-Wesley Longman Publishing Co., Inc. (cited 31 times)	van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworths. (mentioned 71 times)
Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361-371. (cited 29 times)	Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361-371. (mentioned 57 times)
Hill, J. R. (1996). Information seeking in electronic environments. *Journal of the American Society for Information Science, 47*(5), 405-406. (cited 29 times)	Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3-50. (mentioned 53 times)
Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407-424. (cited 28 times)	Marchionini, G. (1997). *Information seeking in electronic environments*: Cambridge University Press. (mentioned 52 times)
Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems, 30*(1-7), 107-117. (cited 25 times)	Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407-424. (mentioned 51 28 times)
Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing & Management, 36*(2), 207-227. (cited	Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6-12. doi: 10.1145/331403.331405 (mentioned 51 times)

23 times)	
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407. (cited 22 times) Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology (ARIST), 29*, 3-48. (Cited 22 times)	Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), 602-615. (mentioned 50 times) Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45*(3), 149-159. (mentioned 50 times)

Table 6 shows that of the top 20 most highly cited works according to CountOne, seven do not appear in the CountX list. When we look at the top 41-60 most highly cited works, we find that there is no commonality whatsoever in the rankings. If we look at these ranks in an accumulative way, the top-ranked works have greater similarity across the two counting methods than the slightly lower ranked works (e.g., the top 1-20 have 65% similarity, the top 1-100 52%). The top 100 ranked works have all have been cited more than once (almost twice) by the articles that reference them. The top 20 ranks by both counting methods have most in common. Examples of significant differences in the two rankings are: “Information needs and uses” ranked 31 by CountOne and 139 by CountX; “User-defined relevance criteria: an exploratory study” ranked 32 by CountOne and 11 by CountX; and “Scholarly communication and bibliometrics” ranked 38 by CountOne and 182 by CountX. There also exist outliers: for example, Seglen's "The skewness of science" is ranked 57 by CountOne and 342 by CountX. This paper was cited and mentioned both 12 times, often in the Introduction section (cited and mentioned 5 times) and the Methodology section (cited and mentioned 3 times). Another extreme is Jorgensen's "Attributes of images in describing tasks" which was ranked 89 by CountOne and 12 by CountX. It was cited 10 times but mentioned 48 times, especially in the Methodology section. Hsin-Liang Chen's *JASIST* paper in 2000 "An analysis of image queries in the field of art history" considers Jorgensen's as one of the major methods to investigate end users' image queries in art history. Also Abebe Rorissa's *JASIST* paper in 2010 "A comparative study of Flickr tags and index terms in a general image collection" has extended Jorgensen's approach to large scale social tagging image description and indexing. Both papers mentioned Jorgensen's paper multiply, especially in the Methodology section.

Table 6: Differences between CountOne and CountX rankings for the 100 most highly cited works (the left three columns are based on the even distribution of top 100 by every 20, and the right three columns are the accumulated ranking of top 100 by adding 20 more ranks each time)

Top-100 most highly cited papers	% of difference (CountOne vs. CountX)	Avg # of mentioned times/paper	Top 100 highly cited papers	% of Difference (CountOne vs. CountX)	Avg # of mentioned times/paper
Top 1-20	7 (35%)	1.939	Top 1-20	7 (35%)	1.939
Top 21-40	11 (55%)	1.987	Top 1-40	13 (32.5%)	1.958
Top 41-60	20 (100%)	1.868	Top 1-60	27 (45%)	1.928
Top 61-80	18 (90%)	1.918	Top 1-80	40 (50%)	1.926

Top 81-100	18 (90%)	1.988	Top 1-100	48 (48%)	1.938

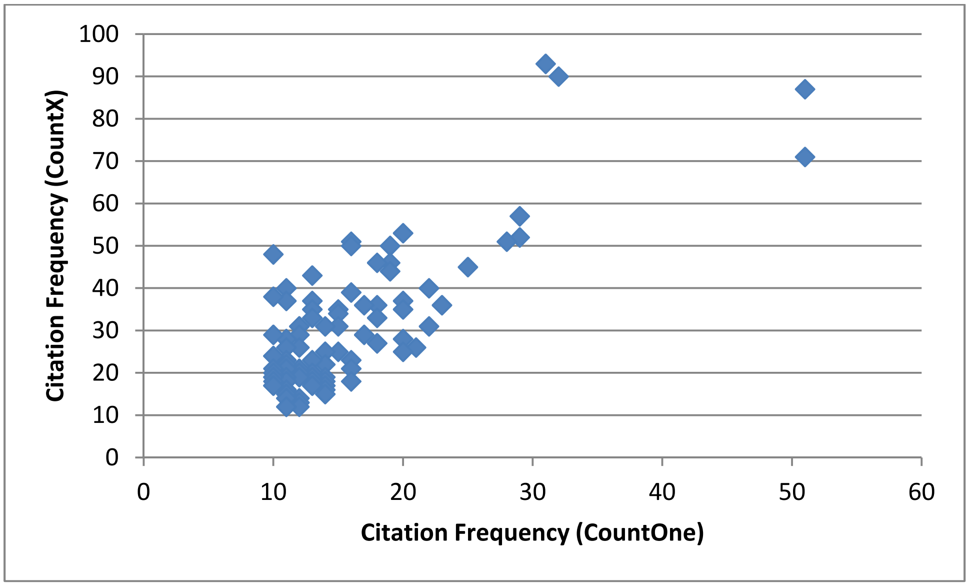

Figure 1. Scatter plot of citation frequency (CountOne vs. CountX)

Figure 1 is a scatter plot of citation frequencies of the top 100 highly cited works using CountOne and CountX. The Pearson product-moment correlation coefficient is 0.779, $p<0.01$ (2-tailed), which reveals a strong correlation between the two counting methods. Figure 2 is a scatter plot of the rankings obtained using CountOne and CountX for the top 100 highly cited papers. A moderate rank order correlation between the two sets of results was obtained: Spearman's $r = 0.589$, $p<0.01$ (2-tailed). The correlation of the rankings between CountOne and CountX is statistically significant ($p<0.01$). The coefficient of determination is the square of r (0.589*0.589=0.347) which explains that 34.7% of the ranking based on CountOne can be explained by the ranking based on CountX, while 63.7% cannot. As Table 6 shows, the two sets of top 20 ranks are quite similar, but they diverge as we move from the top 20 to the top 100 references. Therefore, CountOne (the approach traditionally favored in citation analysis) and CountX (the alternative proposed here) can generate different citation ranks for the same set of bibliographic references.

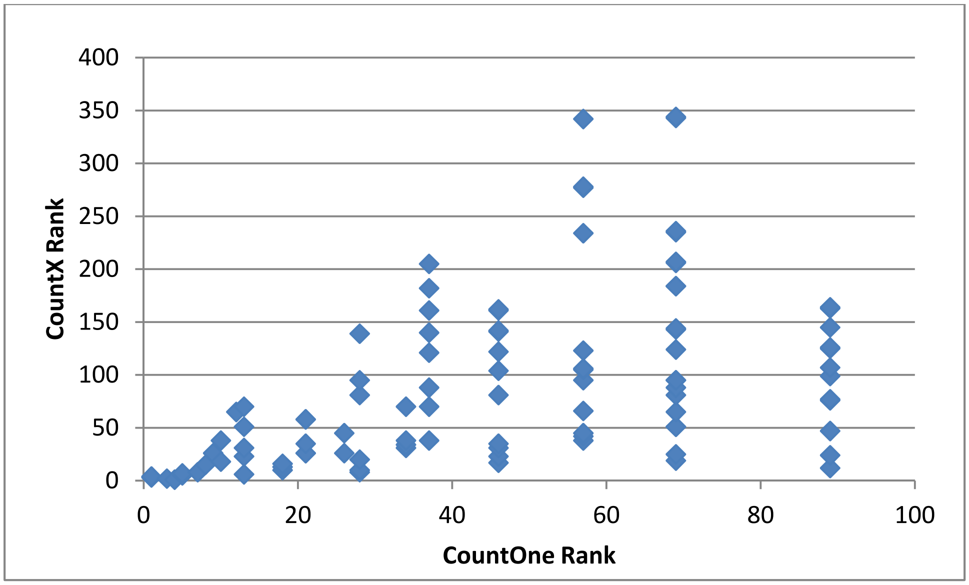

Figure 2: Scatter plot of ranking differences between CountOne and CountX

Conclusion

Voos and Dagaev (1976) suggested that the number of times a reference is cited in a paper provides some indication of its relevance to the citing paper's subject. Indeed, as Small (1987, p. 339) pointed out "there is a great deal of evidence that 'influential' papers are more highly cited than 'uninfluential' ones. There is no evidence to suggest that highly referencing papers are highly 'influential', whatever that might mean. Of course much further work could be done on this topic."

We conducted a content-based citation analysis of research articles published in *JASIST* (2000-2011) in order to explore the extent to which bibliographic references (that is, the works they denote) contribute differentially to the articles that cite them. Our approach involved 1) analyzing the distribution of citations across article sections and 2) counting the number of times each reference was cited/re-cited in the body of the text. We found that most highly (re-)cited works (using CountOne) in our data set tended to appear in the Introduction, Literature Review, and Methodology sections whereas using the CountX method, the top 3 locations for highly cited works are actually the Methodology, Literature Review, and Introduction sections. We also found that rankings of highly cited works were influenced by whether one counted single or multiple mentions of a bibliographic reference in the text. Of course, the architecture of journal articles, the compositional norms and stylistic preferences of a discipline and the referencing behaviors of authors may vary considerably, so what holds for information science may not apply in the case of, say, chemistry or comparative literature. Still, as Herlach (1978) has pointed out, a work that has been cited in the Introduction or Literature Review, and subsequently mentioned in the Methodology or Discussion sections, will likely have made a more significant contribution to the citing article than one which has been mentioned only once in the entire article. In other words, in assessing the contribution of a bibliographic reference we should, ideally, take account of the number of times the work in question is re-cited and the location of those re-citations throughout the citing article.

A few avenues for further research suggest themselves. First, in most cases the text surrounding the citation provides clues to help classify the functions performed by the citations in the citing paper (Peritz, 1983). We thus intend to extend the topic modeling algorithm (e.g., Latent Dirichlet Allocation) to extract topic distributions (Blei, Ng, & Jordan, 2003), or citation sentiments (Small, 2011), from the text surrounding the citations in citing articles. Second, Peritz (1983) discussed the heuristic value of the roles played by citations in the citing paper to assess the nature of the contributions made by a defined body of literature to a field. We plan to evaluate citations based on their contributions to either formulating a research question or providing the fundamental methodology by using machine learning algorithms on large-scale, full-text document collections. Third, citations have been described as "frozen footprints on the landscape of scholarly achievement" (Cronin, 1984, p. 25). Once the footprints have been made in the journal article of record, they can't be updated. The limiting temporal nature of citations should also be considered when conducting content-based citation analysis. Fourth, more data can be collected using surveys of or interviews with selected author groups to identify nuanced features of referencing behavior (e.g., the rationale for mentioning some references once and others multiply, the reasons for placing references in different sections of articles). Furthermore, the marrying of content-based citation analysis with traditional bibliometric approaches could lead to enhanced techniques for mapping scientific disciplines and evaluating research impact (Glenisson, Glänzel, Janssens, & De Moor, 2005).

References

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries*, 68(5), 314-316.

Blei, D., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

Bollen, J., Rodriguez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69(3), 669-687.

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

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? Social Studies of Science, 5, 423-441.

Cronin, B. (1981). Agreement and divergence on referencing practice. Journal of Information Science, 3, 27-33.

Cronin, B. (1984). The citation process: The role and significance of citations in scientific communication. Taylor Graham: London.

Cronin, B. (2004). Normative shaping of scientific practice: the magic of Merton. Scientometrics, 60(1), 41-46.

Ding, Y. (2011). Applying weighted PageRank to author citation networks. Journal of the American Society for Information Science and Technology, 62(2), 236-245.

Finney, B. (1979). The reference characteristics of scientific texts. London: City University, Centre for Information Science (M.Sc. dissertation).

Frost, C. (1979). The use of citations in literary research: Preliminary classification of citation functions. Library Quarterly, 49(4), 399-414.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178(60), 471-479.

Garfield, E. (1977). The citation index as a subject index. Essays of an Information Scientist, 2, 62-64.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. Information Processing & Management, 41(6), 1548-1572.

Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the 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.

Hodges, T. (1978). Citation indexing: Its potential for bibliographical control. Berkeley: University of California (Ph.D. thesis).

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632.

Lipetz, B. A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. American Documentation, 16(2), 81-90.

MacRoberts, M. H. & MacRoberts, B. R. (2010). Problems of citation analysis: A problem of uncited and seldom-cited influences. Journal of the American Society for Information Science and Technology, 61(1), 1-12.

McCain, K., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. Scientometrics, 17(1-2), 127-163.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. Science, 159, 56-63.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. Social Studies of Science, 5, 86-92.

Murugesan, P., & Moravcsik, M. J. (1978). Variations on the nature of citation measures with journals of scientific specialities. Journal of the American Society for Information Science, 29, 141-147.

Oppenheim, C. & Renn, S.P. (1978). Highly cited old papers and the reasons why they continue to be cited. Journal of the American Society for Information Science, 29(5), 225-231.

Peritz, B. C. (1983). A classification of citation roles for the social sciences and related fields. Scientometrics, 5, 303-312.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory with applications to literature of physics. Information Processing & Management.

Small, H. (1976). Structural dynamics of scientific literature. International Classification, 3(2), 67-74.

Small, H. (1978). Cited documents as concept symbols. Social Studies of Science, 8, 327-340.

Small, H. (1982). Citation context analysis. Dervin, B. J., & Voight, M. J. (Eds). Progress in communication sciences. Norwood, NJ: Ablex, Vol. 3, 287-310.

Small, H. (1987). The significance of bibliographic references. Scientometrics, 12(5-6), 339-341.

Small, H. (2011a). Interpreting maps of science using citation context sentiments: A preliminary investigation. Scientometrics, 87, 373-388.

Small, H. (2011b). Text mining of citations for emerging scientific communities. Talk presented at Mapping and Measuring Scientific Output, Los Alamos National Laboratory, May 11. Available online at: http://mediazone.brighttalk.com/comm/ReedElsevier/78d7d945a8-24781-4950-26860

Spiegel-Rosing, I. (1977). Science studies: Bibliometric and content analysis. Social Studies of Science, 7(1), 97-113.

Suppe, F. (1998). The structure of a scientific paper. Philosophy of Science, 65(3), 381-405

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP06), Sydney, Australia, July 22-23, 2006.

Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or Did we op. cit. your idem? Journal of Academic Librarianship, 1, 20-21.

White, H. D. (1990). Author co-citation analysis: Overview and defense. In: Borgman, C. L. (Ed.). Scholarly communication and bibliometrics. Newbury Park, CA: Sage, 84-106.