Semiometrics and Impact Calculations

D. M. McRae-Spencer, N. R. Shadbolt

School of Electronics and Computer Science University of Southampton
E-mail: {dmms03r, nrs}@ecs.soton.ac.uk

Abstract. Citation analysis and journal impact factors are considered controversial yet have been a standard measure of research value for over thirty years. This paper considers the nature of citation, the arguments for and against the current approach to impact measurement based on citation count and proposed alternative measures. It is argued that most proposed alternatives are attempting to do the same thing: apply a value-based approach to the impact measurement, more subtle than purely counting citations. Parallels are drawn with psychology research and in particular the emerging field of Semiometrie, and it is argued that applying the semiometric approach to the area of citation analysis and impact measurement provides not only a method to unify the proposed alternatives, but suggests a future scenario of research measurement and impact analysis far richer than has previously been considered possible.

Why do articles become highly cited?

The practice of scientific authors citing previously published papers has been around as long as scientific publishing itself. The citation graphs thus produced track the genealogy of scientific advances over the centuries, and in today's world of global computer networks and large-scale databases, it is possible to track such graphs on a scale never before possible. Analysis of citation graphs has been used for over thirty years as the primary means of assessing journal impact (Garfield 1972)[1] and citations have become an integral metric in determining scientific impact.

In investigating the science of citation analysis, it is worth firstly asking a question often overlooked in such studies: why do articles get cited at all, and why do certain papers become more highly cited than others? While there are some clearly obvious answers – writers will cite documents that are relevant to their current topic – it is important to review work that has been conducted in this area. Two studies in particular offer large-scale surveys of motives for citation: a psychology-centred study by Shadish et al in 1995[2] and a 2000 paper by Case and Higgins looking at communications studies[3]. Case and Higgins also review the previous work on citation theory and track the shifts between 'normative' theory (citation is due to relevance) and 'persuasional' theory (citation is due to self-interest), arguments that, without sufficient empirical basis, led Cronin (1984) to conclude that "it is difficult to see how citation can be defined as a norm-regulated activity"[4]. Adopting a practise of surveying people who cited particular papers, both studies asked the citers to choose their reason

for citation from a list of 28 (Shadish et al.) or 32 (Case/Higgins) possible reasons, basing those lists on the previous theoretical work. In both cases, responders were allowed to choose more than one reason, but had to give relative importance values for these reasons.

The results were largely consistent between the two studies (Shadish et al. pp.481-2, Case/Higgins p.640). Shadish et al. identified six specific citation types ("exemplar citations, negative citations, supportive citations, creative citations, personally influential citations and citations made for social reasons") while Case and Higgins identified seven ("classic citation, social reasons for citing, negative citation, creative citation, contrasting citation, similarity citation and citation to a review"). Case and Higgins further refined both their own work and that of Shadish et al. by drawing out the three most significant factors for citation: "first, the perception that the work is novel, well-known and represents a genre of studies; second, the citing author's judgment that citing a prestigious work will promote the cognitive authority of his or her own work; and third, the perception that a cited item deserves criticism – which can also serve to establish the citer as an authoritative, critical thinker."

While these studies describe the major reasons for citation, it is also important to consider why certain articles are more highly-cited than others. From a citation-reason perspective, Case and Higgins state that while "we cannot reach definitive conclusions about the nature of highly-cited items" they do identify that highly-cited items are: "very likely to emphasize reviews of the literature on their topic, be cited as 'concept markers', and be authored by widely-recognized authorities in a field of research." In addition to these methods, other empirical studies have found interesting features of highly-cited articles. Two conclusions in particular stand out. Firstly, Lawrence (2001)[5] identifies that articles available online are cited on average 336% more than those not available online. Secondly, and related, McVeigh (2004)[6] notes that Open Access journals are increasing in number such that "over 55% of the article content indexed by Thomson ISI in 2003 was produced by a publisher that allows some form of author-archiving."

What is the standard for impact measurement?

Although these studies are relatively recent, the use of the citation graph analysis is much older. Garfield (1972)[1] proposed that the citation count of a collection of papers (such as a journal) was directly proportional to the scientific importance of that collection, based on his earlier (1955)[7] suggestion that "reference counting could be used to measure impact"[8]. As a result of this, the Institute for Scientific Information (ISI), founded by Garfield, has published an annual Journal Citation Report (JCR) based on the following calculation, termed the 'Impact Factor': "it is a measure of the frequency with which the 'average article' in a journal has been cited in a particular year or period. The annual JCR impact factor is a ratio between citations and recent citable items published. Thus, the impact factor of a journal is calculated by dividing the number of current year citations to the source items published in that journal during the previous two years."[9]

Calculation for journal impact factor, *from Garfield (1994)[9]*
A= total cites in 1992
B= 1992 cites to articles published in 1990-91 (this is a subset of A)
C= number of articles published in 1990-91
D= B/C = 1992 impact factor

Initially impact factors were introduced "primarily as a bibliographic research tool for retrieval of overlapping research for the benefit of scientists who worked in relative isolation to contact colleagues with comparable interests"[10]. However, since these were first published in 1972, the JCR figures have become the standard metric for determining journal impact across sciences and social sciences. Indeed, in terms of research evaluation, journal impact factors are "probably the most used indicator besides a straightforward count of publications"[11].

What are the criticisms of this approach?

While standard, the JCR figures on impact factor are far from perfect. Garfield himself admits that Journal Impact Factors are "controversial" and notes that "the literature is replete with recommendations for corrective factors that should be considered, but in the final analysis subjective peer judgment is essential."[8]. While it is hard to disagree with that statement, many authors have questioned the validity of even using Impact Factor measures at all. In particular, Seglen (1992[12], 1994[13], 1997[11]) notes that there are a number of very good reasons not to use journal impact factors for evaluating research under any circumstances:

- Use of journal impact factors conceals the difference in article citation rates (articles in the most cited half of articles in a journal are cited 10 times as often as the least cited half).
- Journals' impact factors are determined by technicalities unrelated to the scientific quality of their articles.
- Journal impact factors depend on the research field: high impact factors are likely in journals covering large areas of basic research with a rapidly expanding but short lived literature that use many references per article.

(from Seglen 1997[11])

Opthof[10], unlike Seglen, argues that impact factors can be legitimately used to judge the research impact of journals but draws a number of very clear restrictions on the use of this metric, most notably that journal impact factors may not be legitimately applied to individual papers, authors or groups of scientists (such as research groups or institutions) who produce fewer than 100 papers in the JCR-standard two year period of measurement.

Beyond the statistical validity or otherwise of journal impact factors, there are other considerations that need to be taken into account. For instance, it is worth noting that in today's online age, 'citation lag' is shortening and thus the two year standard may not be the correct timescale on which to judge citation impact, although given the

varying frequency of journal publications, it may not be meaningful to reduce the standard to below a two year figure[14]. Additionally, there are arguments against the types of citations used in impact measures. Seglen notes that "self citations are not corrected for"[11] in journal impact factor measurement, leading to self-inflation; while Gabehart[15] points out that articles later retracted by journals are frequently positively cited due to there being no method of tying a retraction item to the original article in citation analysis. All the above contribute to the "controversy"[8] surrounding the use and abuse of journal impact factors, and the debate continues.

What are the alternatives?

There is clearly, therefore, a demand for alternative methods for determining impact of papers, authors, institutions and even journals. While the citation graph is, and will remain, the primary method for determining whether work is relevant ("normative" theory accounts for the largest sub-group of reasons for citation in the studies by both Shadish et al and Case and Higgins), a number of alternative methods have been suggested and applied. Kleinberg quotes a 1976 study by Pinski and Narin[16], noting their "more subtle citation-based measure of standing, stemming from the observation that not all citations are equally important. They argued that a journal is 'influential' if, recursively, it is heavily cited by other influential journals."[14] However, such algorithms are computationally expensive and perhaps unrealistic as a practical alternative to traditional impact factors, at least until recently.

Kleinberg draws connections between Pinski and Narin's work and his own hyperlink analysis algorithm for determining hubs and authorities on the web. While noting that document purpose is different in the two fields, and thus weightings in the algorithms will be different, there are clear parallels in the processes involved. Indeed, the large-scale citation network engine Citeseer[17] "aims to identify hubs and authorities in the scientific literature"[18] by applying Kleinberg's techniques to its current corpus of over 700,000 documents.

In separate work, Chen's CiteSpace[19] application looks to identify 'landmark', 'hub' and 'pivot' nodes, specifically with the purpose of finding Kuhnian[20] turning points in scientific development. By applying visualisation-over-time analysis and adding pruning techniques such as Pathfinder, Chen's work not only offers new methods of visualising scientific progress but also backs up the theoretical work of Kleinberg and Kuhn, and offers up the question that is key to the search for new metrics: "is it possible that an intellectually significant article may not always be the most highly cited?"

Other proposed metrics include the application of PageRank algorithms[21] to citation graphs, download/viewing statistics as a part of the impact factor[22] and the application of acknowledgement analysis as part of author impact calculation[23] – an approach which, when combined with citation indexing, "yields a measurable impact of the efficacy of various individuals as well as government, corporate and university sponsors of scientific work."[23]. With such diverse approaches being suggested and applied, the increase in online availability of documents and the emergence of large-scale citation networks, the question arises: can these approaches be generalised or

even linked such that an integrated approach can be applied to the impact metric problem? This report suggests that semiometrics may be the answer, and it is to this subject that we now turn.

What are semiometrics?

Semiometrie is, specifically, an empirical approach developed by Steiner et al[24] that is used to determine personality type. A carefully-chosen list of 210 words are presented to the subject, who is asked to rate their emotional response to those words on a scale of strongly negative to strongly positive. These results are collated and the personality type – of an individual, a group, a nation – is then determined. Camillo et al. note that this approach "is based on the principle that words are not only significant of things, but they refer to values and affections to which a single or a group of people are related."[25]

In practical usage, the semiometric approach has become increasingly associated with media research. The UK television company Channel 4, for example, commissioned a semiometric study into the types of people who watched various of their programmes, in order to pass on audience demographics to other interested bodies such as advertisers[26]. Their description of semiometrics as "pushing back the boundaries between quantitative and qualitative"[27] is similar to the conclusion made in the text-mining community, where Camillo et al. have used Semiometrie to track the type of people who post on internet sites – tracking their usage (positive or negative) of Steiner et al.'s original list of 210 words. In numerous fields, it seems, the semiometric approach is being increasingly applied to acquire results that are both quantitative and qualitative.

How could Semiometrie be applied to citation analysis?

The question of whether Semiometrie has anything to offer citation analysis at all is a legitimate one. It certainly isn't the case that rating the personality types of document authors according to the 210 word list will help calculate their research impact, nor is it relevant to determine the dominant personalities involved with particular research domains, although it would be an interesting study. The use of semiometrics in citation impact studies is an application of principle rather than straight re-use Steiner's 210-word method. The founding principle of Semiometrie is that ratings and judgements can be made on an object using a negative to positive (numerically, -3 to +3) scale of, in Steiner's case, emotional responses. In the context of citation analysis, this implies an approach of rating citations on, say, a -3 to +3 scale, rather than treating all citations as equally valuable.

However, we are not simply asking "how contextually positive or negative was this citation" (work that again draws on Case/Higgins and Shadish). That question is key to the whole process, but it is important to note that citation rating also asks such questions as "how important is this citation?" – an important citation should be given a higher semiometric score than a less important one. This is exactly what Kleinberg's

authority measurement[14] does: its iterational method is based on the idea that citation by a 'hub' means a document is more likely to be an 'authority'. Thus, any citations by a 'hub' document should be given a higher semiometric score. Kleinberg's approach – and that of Pinski and Narin[16] – can therefore be seen as one example of the semiometric approach.

On a higher level, papers and even authors may be rated in the same way, with the sum of citation rankings comprising only a part of paper rankings, and the sum of paper rankings comprising only a part of author rankings. Therefore other examples of semiometric measurements would include the application of Chen's work[19] in scoring a particular paper with regards to whether it is a hub, a landmark, a pivot point, a turning point and perhaps its geographical centrality to the visualised citation graph – giving a semiometric measure for both a paper itself and for its citations to other papers. The acknowledgements of a given paper, similarly, yield semiometric measures (albeit almost always positive) for a person or institution. In such cases, the final semiometric rating for a person would comprise the summation of their individual paper ratings along with their acknowledgement rating, weighted as appropriate. Internet PageRank rating, download/viewing statistics, citation half-life analysis and, yes, pure citation counting should also apply (every citation may perhaps be given a starting value of 1 on the scale to distinguish it from non-cited papers, and this could rise or fall depending on the other factors).

Beyond graph analysis, two further major sources of information should be used as part of the semiometric analysis for citations and papers. Firstly for citations, natural language text processing techniques could be developed to automatically determine, albeit roughly, the nature of the citation on a negative-positive scale. Secondly for papers, peer review responses should be included in the measures because the single best way to identify an important paper is to ask an expert: as Garfield himself notes, "the literature is replete with recommendations for corrective factors that should be considered, but in the final analysis subjective peer judgment is essential."[8] The above list of possible semiometric measures is not exhaustive, and indeed the identification of a common scale for determining research impact at different levels may lead to the discovery or invention of new methods. However, at this point two things are clear: (1) there is a desire in the scientific community for a realistic alternative to traditional journal impact factors as a measure of research impact, and (2) the semiometric approach allows for the amalgamation of a number of the disparate alternatives that have been presented to the scientific community over recent years.

Conclusion

The application of the semiometric approach to research analysis can been defined as "looking at a whole range of measures and content implicit in the Conventional Web and meta content explicit in the Semantic Web to build much richer views as to the real research activity underway in our discipline."[28] This paper has summarised current literature surrounding the nature of citations and identified a number of citation types, including negative citations. The history of citation analysis as a source of journal impact factors and research ratings has also been discussed, along with a

summary of the dissatisfaction felt by many scientists from diverse fields of this method, dissatisfaction that is often centred around the observation that highly influential articles are not always highly cited. Proposed alternative measures have been discussed and the argument put forward that a number of these measures are semiometric in nature: interested not just in the number of citations, but in the quality of those citations, papers, authors and groups, according to a wide variety of measures, and as such they may be amalgamated into one overall measure. This paper also argues for the inclusion of natural-language processing to determine the contextual nature of citations, and the adoption of peer-review judgements as part of the semiometric analysis. Future work in this area includes not only research into each of these areas, and studies into the relative weightings that will need to be applied to each factor, but also into the application of semiometric analysis to the increasingly-large online paper repositories and archives and the citation graphs they create. The global network of online citations allows algorithmical analysis of papers and research in a manner never before seen, and the value-based semiometric approach will allow research and impact analysis far richer and more subtle than has been possible before.

References

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

2 Shadish, W., Tolliver, D., Gray, M., & Gupta, S. "Author Judgements about works they cite: Three studies from psychology journals." Social Studies of Science, 25, (1995). 477-497.

3 Case, D.O., Higgins, G.M. "How can we investigate citing behaviour? A study of reasons for citing literature in communication." Journal of the American Society for Information Science, 51(7), (2000). 635-645.

4 Cronin, B., The citation process. London: Taylor Graham. (1984).

5 Lawrence, S., "Online or invisible." Nature, 411(6837), (2001). 521.

6 McVeigh, M.E., "Open Access Journals in the ISI Citation Databases: Analysis of Impact Factors and Citation Patterns." Thomson Corporation. (2004).

7 Garfield, E., "Citation indexes for science: a new dimension in documentation through association of ideas." Science, 122, (1955). 108-111.

8 Garfield, E., "Fortnightly Review: How can impact factors be improved?" British Medical Journal, 313, (1996).411-413.

9 Garfield, E., "The Impact Factor." Current Contents, Thomson Corporation, June 20 edition, (1994).

10 Opthof, T., "Sense and nonsense about the impact factor." Cardiovascular Research, 33, (1997). 1-7.

11 Seglen, P.O., "Why the impact factor of journals should not be used for evaluating research." British Medical Journal, 314, (1997). 497.

12 Seglen, P.O., "The skewness of science." Journal of the American Society for Information Science, 43, (1992). 628-38.

13 Seglen, P.O., "Causal relationship between article citedness and journal impact." Journal of the American Society for Information Science, 45, (1994). 1-11.

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

15 Gabehart, M.E., “An analysis of citations to retracted articles in the scientific literature.” Master’s Paper for the M.S. in L.S degree, University of North Carolina. (2005).
16 Pinski, G., Narin, F., “Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics.” Information Processing and Management, 12, (1976). 297-312.
17 Lawrence, S., Bollacker, K., Giles, C.L., "Digital Libraries and Autonomous Citation Indexing" IEEE Computer, 32(6), (1999). 67-71.
18 Lawrence, S., Bollacker, K., Giles, C.L., “Indexing and Retrieval of Scientific Literature." Eighth International Conference on Information and Knowledge Management, CIKM 99, (1999). 139-146.
19 Chen, C., “Searching for intellectual turning points: Progressive knowledge domain visualization.” Proceedings of the National Academy of Sciences, 101(Suppl. 1), (2004). 5303-5310.
20 Kuhn, T.S., The structure of scientific revolutions. 2 ed., Chicago: University of Chicago Press. (1970).
21 Page, L., Brin, S., Motwani, R., Winograd, T., “The PageRank Citation Ranking: Bringing Order to the Web.” Stanford Digital Library Technologies Project. (1998).
22 Bollen, J., Van de Sompel, H., Smith, J.A., Luce, R., “Toward alternative metrics of journal impact: A comparison of download and citation data.” Information Processing and Management, Special issue on Informetrics. (2005).
23 Councill, I.G., Giles, C.L., “Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing.” Proceedings of the National Academy of Sciences, 101(51), (2004). 17599-17604.
24 Steiner, J.F., Lebart, L., Piron, M., La sèmiométrie. Dunod Parigi. (2003).
25 Camillo, F., Tosi, M., Traldi, T., “Semiometric approach, qualitative research and text mining techniques for modelling the material culture of happiness.” Thinking creatively in turbulent times, Bethesda, Maryland: World Future Society. (2004).
26 “Understanding the subconscious desires of Channel 4 viewers.” TNS Media/Channel 4: (2004).
27 http://www.in4mer.com/research_viper.asp?section=semiometrie: The In4mer Research Website. (2004).
28 schraefel, m.c., Shadbolt, N.R., Gibbins, N.M., Harris, S.W., Glaser, H., “CS AKTive space: representing computer science in the semantic web.” Proceedings of the 13th international conference on World Wide Web, (2004). 384-392.