Research Trends

Volume 1
Issue 37 *Special issue on alternative metrics*

Article 3

6-1-2014

Predicting citation counts

Ron Daniel Dr
Elsevier Labs

Section 2:
Behind the data

Predicting citation counts

Ron Daniel Jr., Ph.D.
Director, Elsevier Labs

Abstract

Many articles have been written about efforts to predict how many citations a research article will receive, based on indicators available before or shortly after publication. These efforts have widely varying results, with one effort predicting 14% of the variance in citation while another study ten years later reached over 92%. What was learned in that decade? What can this tell us about potentially valuable altmetrics, and are there areas in which new altmetrics might be discovered?

1. Introduction

This special issue of Research Trends is about altmetrics – alternatives to the use of citation counts as the metric for assessing the impact of an article, a researcher or a journal. Citation counts do not tell the whole story (e.g. they don't value useful research software tools, useful advisory papers to young researchers, or research that can't be published for commercial or government security reasons). Having additional metrics to provide a more complete picture is a very welcome development. However, even in a future in which additional metrics are available to assess impact, citation counts will remain first among equals because of their intimate connection with the text of the article and the article's basis on prior work.

The current and continued importance of citation counts has led to the desire to predict how often an article will be cited in order to predict its future importance. Such predictions could be used to decide if an article should be published in one journal vs. another, to flag new research for scrutiny before citation counts have had time to accrue, to assess the development of a young researcher before many counts could have accrued, etc. Many articles have been written on this topic, but there has been very little consistency in their results. Four studies between 2002 and 2012 found that they could predict 14% (1), 20% (2), 60% (3) and 90% (4) of the variance in citation counts a few years after publication based on features available before or shortly after publication. A discrepancy this great requires some explanation!

This article has three goals. The first is to explain the discrepancy in the previous research results. The second is to evaluate the various indicators, a.k.a. features, which were used in the four articles. Features that are predictive of eventual citation counts might be particularly valuable altmetrics that serve as leading indicators of an article's merit. We need to be cautious when comparing results across the studies as they use different scientific domains, make predictions over different time periods, use different statistical methods, obtain results through different procedures, etc. For example, one study measured "newsworthiness" by having readers estimate it; another did so by searching news archives. Both found it to be a notable factor but not necessarily statistically significant. All of this means the results are only loosely comparable. However, we can look within each study to see which features did have significant effects and the relative magnitude of those effects. If the same feature is found to be significant, or not, across all studies then we are fairly safe in drawing conclusions about its utility. The third goal is to see if we can draw conclusions about potentially valuable altmetrics and areas where new altmetrics might be discovered.

2. Prediction Features for Pre-Publication Articles

From the time an article is first conceived, features begin to accrue that we can use to predict its future citation counts. This section looks at features available from the inception of the article up to the time where it has been accepted for publication in a journal. The features can be further subdivided into those that apply to the article itself, to the authors of the article, and to the journal which has accepted the article for publication. Those three categories are named as Content, Author, and Venue (4). What we will see is that even before an article is published, we have enough information to make fairly good predictions about its future rate of citations.

2.1 Content

Study Design Factors:

The earliest article we review (1), published in 2002, made the assumption that high quality research would be more heavily cited. They thought about what made high quality research and looked for corresponding features such as sample size, controls, blinding, etc. Sample size and the presence of a control group were found to have some effect, but not to the level of statistical significance. The other factors (blinded, randomized, prospective v. retrospective) were even weaker. The second article (2), published in 2007, also looked at study design factors and found them to have little effect. What they did find, however, was that large studies funded by industry, and with industry-favoring results, were statistically significant predictors of higher rates of citation in the future. These features are understandably important in the medical therapeutic space. Such studies are likely to show drugs and other therapies soon to be available. These factors don't seem likely to generalize to other domains.

Topic:

Unlike the first study, which was confined to emergency medicine, the second study (2) considered the effect of the topic of the article. They found that cardiovascular and oncology articles were more likely to be cited than those on other topics such as anesthesiology, dermatology, endocrinology, gastroenterology, etc. Given the relative death rates of heart disease and cancer to the implications of the other specialties, this seems reasonable. Similarly, the third article (3), published in 2008, found that articles which provided therapeutic information were more cited, as were those which provided original research as opposed to review articles. That study also found that longer articles were cited fewer times, in a weak but statistically significant way. It also found that the more references an article contained, the more likely it was to be cited, although this effect was weak and not significant. The fourth article (4), published in 2012, found a weak effect that the more topics an article covered the higher the number of citations it received.

Table 1 lists the content-based features available before publication which were used in the four studies. Statistically significant values are highlighted. The key things to notice in this table are how few content-based features are significant, and how few of the features are used in multiple studies.

	Callaham 2002 (1)	Kulkarni 2007 (2)	Lokker 2008 (3)	Yan 2012 (4)
# study participants	26.5%	3.1, p=.04	< .001, p=.295	
Newsworthiness score	26%	13.5, p<.001	.133, p=.161	
Control group	24.3%			
Quality score	15.8%			
Explicit hypothesis	4.7%			
Prospective v. retrospective study	2.7%	3.6, p=.01	.477, p=.009	
Type of study participants	2.1%			
Blinded	.07%			
Randomized	0	13.4, p=.01		
Positive results	0			
Industry funding		19.9, p<.001		
Industry favoring result		19.4, p<.001		
Location of study		11.9, p=.001		
Topic		17.8, p=.001		
Original v. review article			.477, p=.009	
# pages			-.011, p< .001	
Structured abstract			-.8, p=.002	
# cited references			.004, p=.008	
Multicenter study			.367, p=.014	
Therapy v. other article			.339, p=.023	
Word count of abstract			-.0003, p=.658	
Semi-structured abstract			.071, p=.746	
Nation of first author			-.037, p=.762	
Novelty				.059
Topic rank				.079
Diversity of topics in article				.157

Table 1: Content-based features available pre-publication.

2.2 Author

The effects of the author were not considered in (1). The second study (2) only looked at whether the author byline indicated group authorship. This was found to be the most significant prediction feature in their study! This was a very important result. It indicated that article importance or quality was not easily measured by the presence or absence of some features we might call "good research practice". That realization led to significantly improved prediction accuracy in later work.

The last two papers (3, 4) looked at author-related features in more detail.

Both Lokker (3) and Yan (4) looked at the count of the number of co-authors. Lokker (3) found that count to be a significant factor, but Yan (4) did not. Yan looked at several other author-related features. The Maximum Past Influence of the Author (MPIA) is the citation count for the author's most-cited paper. The Total Past Influence of the Author (TPIA) is the sum of the citation count across the author's body of work. The MPIA was found to be predictive but the TPIA was essentially useless.

A strong result in (4) was the author's rank in citation counts. The citation counts for all the author's works were averaged, and the average counts were sorted to rank the authors. Figure 1, reproduced from (4), shows that being a very highly cited author is predictive of future citation counts. The rich get richer in other words. As can be seen however, this effect is limited and is only strong for authors in the top ranks of citation frequency.

Considerable attention has been paid to author-related factors in articles beyond the four we review here. (3) provides citations of articles that look at other effects such as nationality, gender, and alphabetic order of the author names.

Table 2 summarizes the effect of the author-based features available before publication. The key thing to notice is that the earliest study made no use of author information, while the latest and most accurate article tried many author-based features.

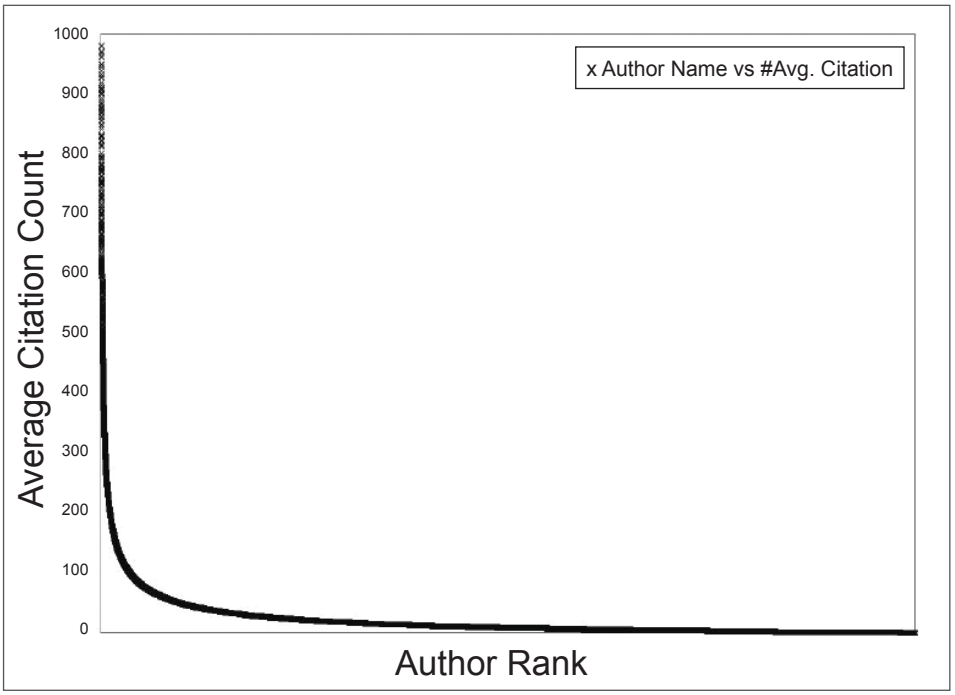

Figure 1: Citation Counts vs. Rank of Author's Average Citation Count. Figure reproduced from Yan et al (2012) (4). We have sought permission for re-use of this figure.

	Callaham 2002 (1)	Kulkarni 2007 (2)	Lokker 2008 (3)	Yan 2012 (4)
# authors		20.3, p<.001	.087, p<.001	.056
Nation of first author			.037, p=.762	
Author rank (by citations)				.593
h-index				.244
MPIA (Max Past Influence)				.585
TPIA (Total Past Influence)				.048
Productivity				.198
Sociality				.249
Authority				.155
Versatility				.160
Recency				.101

Table 2: Author-based features available pre-publication.

2.3 Venue:

The only statistically significant variable found in the first study (1) was the impact factor of the journal in which the article was published. This was an early indication of the power of the venue in determining future citation counts. If we know the journal the article will be published in, we can make more confident predictions about its eventual citation count.

The second study only considered three of the top-line medical journals – JAMA, NEJM, and The Lancet. Nevertheless, they found a significant difference in citation rates between articles in those publications.

The third study did not use the impact factor, as it did not apply to all their sources for content. They discovered other measures that also reflected the article's venue. The strongest are the number of databases that index the journal, and the proportion of articles from the journal which are abstracted within two months by Abstracting & Indexing services and synoptic journals.

Table 3 summarizes the effect of the venue-based features available before publication. Note that no feature is used in more than one study. Curiously, impact factor was the only significant feature found in (1), but it is not used in the later studies. Perhaps the most surprising outcome summarized in this table is the strong effect due to the venues chosen by secondary publication sources like databases, A&I services, and synoptic journals. Given the concerns we all have about infoglut, it is both interesting to see the strength of this effect, and concerning that these effects do not seem to have been featured in any previous altmetric studies. More research in this direction seems justified.

	Callaham 2002 (1)	Kulkarni 2007 (2)	Lokker 2008 (3)	Yan 2012 (4)
Impact factor of publishing journal	Strongest factor, relative contribution = 100%			
Accepted for presentation at meeting	5.5%			
Journal		16.3, p<.001		
Month of publication		0.7, p=.5		
Proportion of articles abstracted			8.18, p<.001	
# databases indexing			.039, p<.001	
Venue rank				.337
Venue centrality				.049
Max past influence of venue				.329
Total past influence of venue				.023

Table 3: Venue-based features available pre-publication.

3. Prediction Features for Newly Published Articles

By publication time, we know many facts about the Content, Author, and Venue. In the newly published phase of the article's lifecycle we shift our attention to early perceptions of the quality of the article, and to early indications of the use of the article.

The previous section showed that venues whose articles were frequently selected for abstraction tended to have more highly cited articles. For a single article, the number of times it is abstracted is also a statistically significant predictor (3) which is not available until shortly after publication. That study also showed that articles which were judged "clinically relevant" by the staff of a recommendation service were significantly more likely to have more citations in the future. These results are notable for the same reason as the venue results in the previous section – secondary publication sources have a predictive effect which is not being captured in current altmetrics.

There are many features that could give us early indications of how often articles are being used, or the perceptions that the early users have of them. Those include:

- Preprint access counts from arXive, etc.
- General Social Media mentions (Twitter, Facebook, …)
- Scientific Social Media mentions (Mendeley, del.icio.us, CiteULike, …)
- Sentiments expressed in early mentions
- Early download counts from services like ScienceDirect
- Early citations of the article shown in services like Scopus

	Callaham 2002 (1)	Kulkarni 2007 (2)	Lokker 2008 (3)	Yan 2012 (4)
Newsworthiness score	26%	13.5, p<.001	.133, p=.161	
Abstracted in evidence based medicine journals			.839, p<.001	
Clinical relevance score			.418, p<.001	
# disciplines rating the article			.038, p=.371	
Time to article being rated			-.009, p=.513	
# views or alerts sent			-.069, p=.938	

Table 4: Features available in first months of publication.

These features were not used in the four studies, but there is good reason to believe that these features will be useful in predicting future citation counts. As mentioned in (3):

"Thirty three percent of the variance in citation counts of BMJ articles were found to be based on counts of online hits and number of pages (5)".

Table 4 shows the effect of features available shortly after the article is published. The most noticeable aspect of this table is that very few post-publication features were used in the studies other than (3).

4. Prediction Features for Mature Articles

The fourth article (4) looked at temporal factors such as age of the article, as well as regression constants to control the growth and decay of citation rates over time. These results were not strong and other studies did not look at features for mature articles so a summary table is not provided. While none of the studies made significant use of features that become available later in the publication lifecycle, there is no shortage of possibilities. For example, we might look at a Page-Rank like scoring of the influence of the papers citing the particular paper of interest.

Nevertheless, the short story is clear. By the time an article is a few months old, we can make good predictions of its likelihood of future citations - especially for those articles which end up being highly cited. Lokker noted that for the papers with the highest citation counts at two years after publication, "Cited articles in the top half and top third were predicted with 83% and 61% sensitivity and 72% and 82% specificity" (3). In other words, only about 20% of the papers which ended up being highly cited were not predicted to be that way.

5. Conclusion

Despite low performance in early studies (14% in 2002), it has become clear over time that it is possible to make good predictions (92% in 2012) of the frequency of future citations. How was this advance achieved? Quite simply, the features being used in the later studies are very different from those used in the earliest ones. The early studies tried to use features around the content, but later work found those to be the weakest while features around the Author and Venue were the most predictive. If we set the power of the Author features to 1.0, the relative power of the Venue and Content features would be about .63 and .25, respectively. We cannot directly compare results across columns, and it is not safe to predict the accuracy any new study might achieve. All of the studies used different domains of literature, predictions over different time periods, different statistical measures, etc. Nevertheless, the pattern seems clear.

It is also interesting, and mildly reassuring, to see that the strongest of these measures operate, to some degree, in a manner independent of each other. Author and Venue are the two most predictive features. However, selecting an article for a journal is usually done in a peer review process that is blind to the identity of the author. Note that this also means these measures are not well-suited for an editorial board to choose articles, since the Venue would be constant and they could not look at the author's publication rank.

In a perfect world, the content of an article would determine its future citation count. We do not, however, have any easily-computed metric for the intrinsic quality and merit of an article. This is where Lokker's results about the importance of secondary sources such as the databases and synoptic journals are most interesting. We see that in the absence of reliable, easily-computed metrics, the subjective human-in-the-loop procedures of peer review, editorial boards, selection for secondary publications, and scientific reputation provide existing mechanisms which fill that void. This provides a potential area of altmetric research to obtain such measures in various fields and compare them with current altmetrics for a variety of purposes.

References:

1. Callaham, M., Wears, R.L., Weber, E. (2002) "Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals", JAMA, Vol. 287, pp.2847-50, Available at http://www.ncbi.nlm.nih.gov/pubmed/12038930/

2. Kulkarni, A.V., Busse, J.W., Shams, I. (2007) "Characteristics associated with citation rate of the medical literature", PLOS ONE; 2:e403, Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0000403

3. Lokker, C., McKibbon, K.A., McKinlay, R.J., Wilczynski, N.L. and Haynes, R.B. (2008) "Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study", BMJ, Mar 22, 2008; Vol. 336, No. 7645, pp. 655–657. Available online: http://dx.doi.org/10.1136/bmj.39482.526713.BE

4. Yan, R., Huang, C., Tang, J., Zhang, Y., and Li, X. (2012) "To Better Stand on the Shoulder of Giants", JCDL, Available at: http://keg.cs.tsinghua.edu.cn/jietang/publications/JCDL12-Yan-et-al-To-Better-Stand-on-the-Shoulder-of-Giants.pdf

5. Perneger, T.V. (2004) "Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ", BMJ, Vol. 329, No. 7465, pp. 546-547, Available at: http://www.bmj.com/content/329/7465/546