Exploring Relationships between Article Level Metrics and Content Quantities in Academic Papers

Arlene J. Casey, Samad Ahmadi, Fionn Murtagh

Abstract Academic papers are growing at the rate of around 3% per year. Navigating this increasing information is challenging for a researcher seeking out high quality content. Our research explores relationships between traditional article level metrics, particularly citations, altmetrics and content quantities, to determine if relationships exist. Understanding these relationships may help to provide indicators to readers about article content prior to reading a paper and guide in paper selection. The research shows that there are positive strong relationships between citations and Mendeley readership counts but not between the citations and the Altmetric score or Twitter counts. This suggests that they are not related or other factors need to be taken into account when looking at altmetrics. We suggest that one of the factors that needs to be accounted for is popular subjects and further exploration is needed to understand how this influences altmetrics. The relationships between the quantities and article metrics are slight but more pronounced when we look at a subset of data with lower Altmetric scores and Twitter counts. There is a relationship between the num-

Arlene J. Casey
De Montfort University, Leicester U.K.,
✉ arlenejanecasey@gmail.com

Samad Ahmadi
De Montfort University, Leicester U.K.,
✉ sahmadi@dmu.ac.uk

Fionn Murtagh
Derby University, Goldsmiths University of London, Derby/London U.K.
✉ fmurtagh@acm.org

ARCHIVES OF DATA SCIENCE (ONLINE FIRST) DOI 10.5445/KSP/1000058749/06
KIT SCIENTIFIC PUBLISHING ISSN 2363-9881
Vol. 2, No. 1, 2017

ber of references within an article and the way in which the citations are used within the article body. The articles studied are retrieved from PLoS (Public Library of Science) with the phrase *text mining*, and thus related content, in their subject or body.

1 Introduction

There are almost 2.5 million articles published a year (Plume and Weijen, 2014) and around 28,100 English language journals listed in Ulrich's directory in 2014 (Ware and Mabe, 2015). They state that article growth has increased on average 3% per annum in 2002 to 2014 and at the same time the number of peer reviewed journals being published has increased at a rate of 3.5% per year. In addition to the growth of papers, technology has made it easier for a reader to access information and search functionality has changed the focus from the traditional journal based to article focused. Many publishers now provide keyword and topic searches and can retrieve the citing papers or similar content papers. The issue that faces the reader is that until they have read the paper they cannot judge the contribution of the paper. Prior to reading, inferences can only be drawn from metrics or knowledge of the author's previous work. Having an understanding of how article level metrics relate to each other and to measures of content could potentially provide an external guidance to the contribution of the paper.

In this paper we will review methods, metrics and altmetrics to measure scholarly impact, the known issues with these and how this may impact their use in indicating scholarly output. We will consider research into content patterns and quantities within academic papers and how this might help us choose relevant quantities to study and compare to determine if relationships exist. We will then explain our data set retrieval, our methodology and experiments run. Finally we present our results and conclusions.

2 Related Work Measuring Scholarly Output

Scholarly output has been tradtionally measured by such metrics as journal impact factors, citations and the h-index, but peer review, the process whereby experts in a research field scrutinise the author's work, is considered the gold standard measure of assessing scholarly output (Ware, 2013). Citations quantify the usage of a scholarly work and are taken as a measure of the research impact. These counts are dependent on the size of a discipline and how many people work within that field. There has been a discussion that some authors and research teams carry out unnecessary self-citations to increase their own citations (Glänzel et al, 2006) and that as co-authorship geolocation increases so do citations (Nomaler et al, 2013). The journal impact Factor (JIF) is a measure of the average number of citations to recent articles in that journal (Garfield, 1998). There are many known issues with JIF such as some journal article content having a higher than average shelf life which impacts the JIF (Whitehouse, 2001). Seglen (1992) has shown that the distribution of citations are skewed with somewhere in the region of 10% of articles receiving over 90% of the citations and that often there is no correlation between the JIF and an individual article within the journal. JIF is also impacted by the size of a discipline and how general a journal may be. There has been research into the potential bias of publishing journals to use self-citations and prefer citations from citing articles within the journal (Ha et al, 2006). The h-index, first suggested by Hirsch (2005), is an author level metric that reflects the productivity of an author or a group of authors. It is the number n of a researcher's papers that have received at least n citations.

2.1 Altmetrics

A number of new metrics have appeared in recent years and these are measures of societal impact of an article based on social web activities. These metrics are known as altmetrics. It is thought altmetrics can provide a timely and responsive gauge of research as they can be collected very quickly. One of the problems with these types of altmetrics is that they are not actually a direct measure of quality or content. Social comments could be positive or negative or nothing about the content e.g. it could be about a funny title. The volume of comments could be impacted by the popularity and size of a subject and by the researchers themselves and their tendencies to use social media. One of the problems when looking at altmetric data is that as more and more people use social media this could lead to incomparable data between different time periods.

A number of studies have looked at the relationship between citations and web based altmetrics. Gunther (2011) showed a high correlation between the number of tweets an article received and the likelihood it would be more highly cited. Costas et al (2014) looked at the altmetric indicators provided by Altmetric.com (Altmetric, 2015) and were able to show a positive but weak correlation between citations and these altmetrics. Haustein et al (2015) studied article metrics and document characteristics across multiple disciplines. They found that citations and social media metrics increased as collaboration increased and as the length of references increased. They suggest though that the factors driving citations and social media are not the same and that they should be seen as complements and not alternatives.

3 Language in Academic Writing

There have been many studies into the use of linguistics in academic writing. One such area is called English for academic purposes (EAP) and focuses on the teaching of academic writing to those who are non-native English speakers. Studies have looked at frequent academic word lists (Vongpumivitch et al, 2009; Coxhead, 2000), verb pattern usage (Zhang, 2015) and the use of nominalisations, where verbs are converted into nouns, within academic text. Through our own teaching work we observe that novice researchers overuse citations without clear purpose, other than what seems to be to increase citation counts. Donohue and Erling (2012) studied if academic attainment could be correlated with a student's use of English for academic purposes. They found there was a strong correlation but that this was associated only with the use of the students' source material rather than other categories of structure, development of text, style and grammar. This suggests the process of turning reading into writing could possibly be an area to consider when defining quantities.

4 Data

The data set comprises scientific papers taken from Public Library of Science (PLoS)(PLoS, 2015). PLoS is a nonprofit publisher that provides open access to its document repository which consists of a suite of journals in science and medicine. All PLoS articles feature article level metrics which include scholarly literature citations, measures of online usage and social comments or bookmarks. We retrieved a document data set of 1200 academic papers with the following criteria: Subject contains *text mining* OR Body contains *text mining* OR semantic similarity, in this case that means the introduction, results and discussion, methodology sections and with a publication date between mid July 2011 and 2014. The search criteria were designed to obtain a data set across three journals from PLoS: PLoS One, PLoS Computational Biology and PLoS Genetics. These criteria are designed to obtain a dataset with some diversity, but largely within a related topic field, but also limited in terms of journal impact differences. For article level metrics we retrieve citations for all our articles and counts for Twitter and Mendeley readership along with their Altmetric score. Citation counts are provided by PLoS who retrieve the values from the relevant third parties in this case Scopus, CrossRef and PMC. We believe this is a representation of traditional measures of citations and newer measures of altmetrics for comparison. All the articles from PLoS are peer reviewed prior to publishing and as such all our articles should be of a minimum standard. We did not consider doing any weightings for co-authorship numbers or self-citations in our analysis. The Altmetric score was retrieved for each paper using the Altmetric API. The Altmetric score is a quantitative measure of how much attention an article has received. The Altmetric score is based on a proprietary algorithm developed by the company Altmetric (Altmetric, 2015).

The data was filtered to remove any potential outliers or problem articles. Where there was no Altmetric score, the data row was removed i.e. no Altmetric score existed as opposed to zero score. Then the range of PLoS altmetric counts were examined and there were several scores that were extremely high. These individual papers were reviewed to understand, if a problem had occurred in pre-processing or if values were genuine. This left 910 papers in the data set. Further examination of the altmetric values was undertaken to assess if the altmetric data set collected retained enough information to ensure the validity of any analysis. From the document collection the % of papers with non zero values were: Scopus – 68%, CrossRef – 67%, PMC – 80%, Mendeley – 80%, Twitter – 67%, Altmetric score – 95%. There were no counts for Facebook entries for any document and this column was subsequently removed. 181 documents had zero for Scopus, CrossRef and PMC with the majority of those having zero for Mendeley. Of those 181 documents only 3% had an Altmetric score of over 10, 4 documents had a score of over 100 and these had high Twitter counts.

Table 1 Data Attributes

Attribute	Notes
Total number of sentences	
Total number of times citations occur	
Number of sentences with a citation	
Average words per sentence	
Number of linked sentences	Sentences following a citation sentence starting e.g. their work, their method.
Repeated citations	How many citations occur more than once.
Multiple repeated citations	How many citations occur more than twice.
Missing citations	References that are not cited.
Number of sentences with multiple citations	
Reference count	Total number of references in bibliography.
PLoS article metrics	
Altmetric score	

4.1 Methods and Experiments

An academic paper represents a written collection of thoughts, details and experiments of an author's work. The general format is comprised of the presentation of research territory, the problem as the author sees it, a discussion of previous related work with a comparison to the author's own work and how their work fits in the wider context. Authors will discuss their experiments, results and finally bring their work to a conclusion. Our premise is that a study of content patterns along with article level metrics may lead to a set of distinguishing features between sets of papers. We believe that these features may indicate the difference between poor or good papers. We study content patterns and investigate any correlations between these and article level metrics for papers.

The content features we consider are listed in Table 1. From the previous discussion it was noted that turning reading into writing is an important aspect of academic writing. Transforming reading into writing is reflected in how an author uses citations and the relevant discussion of the citation work. We therefore focus our content measures on citation information. We look at how citations are made and if they are repeated within the body of an article. We attempt to consider the amount of content related to a citation by marking when a citation occurs and if the subsequent sentence is linked to discussing that citation. We believe that the citation content and how citations are used will provide pattern indications that can distinguish sets of papers. In addition to this we also take a measure of the length of an article by using the total words and average lengths of sentences.

We use correlation to examine the relationship between our data attributes. Bornmann (2014) lists studies that were undertaken looking at correlation of citations and altmetric data. Given the number of studies that use Spearman, we did compare Spearman and Kendall results and found that the correlation values were lower using Kendall, magnitude of approximately 0.05. When looking at correlation we looked at the overall data set of 910 papers and then looked at two smaller subsets of data created by taking all Altmetric scores of 0.25 and under and all Altmetric scores of 50 and above.

5 Results

It is known that typically the distribution of citation data is highly skewed (Seglen, 1992). Altmetrics data has also been shown not to follow a normal distribution (Liu et al, 2013). Our data follows this pattern with histogram plots for all the altmetric counts being highly skewed. These have been log-transformed to enable the distribution to be seen more clearly, see Fig. 1. As an example of the transformation, the original Altmetric data ranged from 0 to 1327.942 and transformed it now ranges from 0 to 7.1921. The Y axis coordinates represent

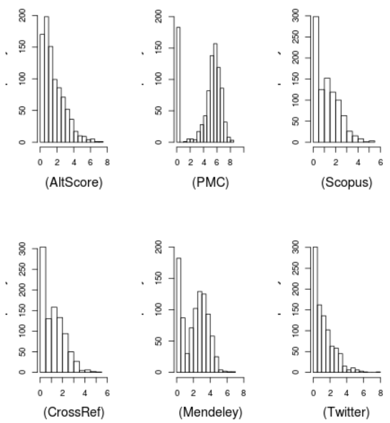

Fig. 1 Histogram plots of the log(x+1)

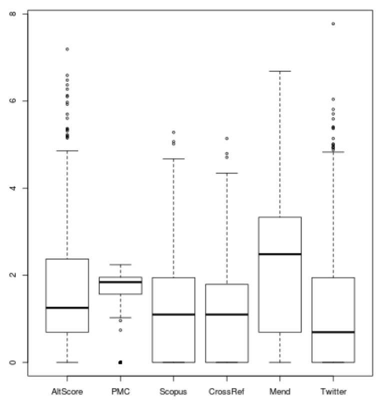

Fig. 2 BoxPlots of article level metrics

the counts or frequencies at each value. The Altmetric Score, Twitter, Scopus Citations and CrossRef Citations are skewed with the distribution mass being on the left of the figure confirming they are right skewed. PMC citations and Mendeley downloads have two peaks that would follow a bimodal distribution. Liu et al (2013) also observe this bimodal distribution for download data but not citations. Looking at the data for PMC counts about 20% of the data is represented in the initial peak, which is data with zero or very low counts. The remaining 80% of the data peaks again with higher citation values. This is quite different to the Scopus or CrossRef citations where the peak is at zero counts and this gradually reduces. This could be a reflection on the readers of PLoS papers and where they publish and thus generate citations. It could also be a reflection on the type of papers that our search extracted with more focus within medical or biotechnology fields that are more likely to get PMC citations.

The boxplots in Fig. 2, for each of the article level metrics, lets us compare between the different metrics and consider the outliers. Altmetric score, Scopus, CrossRef and Twitter all have counts for articles that are very high and these values could be classed as outliers. All of these specific articles were examined and they formed a selection of articles that were either from very popular subjects or were articles based on content around social media use. Where there were high Twitter counts these also had corresponding high Altmetric Scores. Of the three citation metrics PMC looks markedly different to Scopus and CrossRef with more concentrated values and no outliers. As noted above PMC follows a bimodal distribution compared to Scopus and CrossRef. Scopus and CrossRef spread looks very similar as does their distribution in Fig. 1.

Kendall correlation was performed using R. Correlation is considered strongly positive for values 0.4 and above, moderately positive for values between 0.3 and < 0.4, weakly positive 0.15 and < 0.3. Values that are below 0.15 are noted but are probably negligible in terms of a relationship. Unless otherwise indicated the values are significant at the 0.01 level.

Looking at the full data set all citation measures show a strong positive correlation with each other. Mendeley also correlates strongly with all other citation counts, as shown in Table 2. This has been discussed in previous research (Zahedi et al, 2014) where it is suggested that Mendeley readership is a reflection of the numbers of papers that will be cited, the thought being that people do not download and read papers they are not going to then use. The Twitter counts are positive with the citation counts and Mendeley but not to values that are indicative of strong relationships. The Altmetric score and Twitter correlate strongly perhaps an indication about the weighting within the Altmetric.com

Table 2 Kendall Correlation for full data set

	Scopus	CrossRef	PMC	Mendeley	Twitter	Altmetric Score
Scopus	-	0.79	0.62	0.64	0.13	-
Cross Ref			0.6	0.6	0.14	-
PMC				0.61	0.2	0.1
Mendeley					0.2	0.1
Twitter						0.54

Table 3 Kendall Correlation for full data set

	Scopus	CrossRef	PMC	Mendeley	Twitter	Altmetric Score
Number of Sentences	0.13	0.13	0.12	0.14	(-0.02)	(-0.04)
Total number of Words	0.13	0.14	0.12	0.14	(-0.02)	(-0.02)
Sentences with Citations	0.12	0.14	0.11	(0.09)	(-0.05)	(-0.02)

algorithm. The Altmetric score cannot be said to display a strong relationship with the citation counts. This shows that the basis for the Altmetric score is influenced more by Twitter than by citations.

The correlations between the content quantities and the altmetrics for the full data set are most likely negligible but nonetheless it is worth considering which quantities have possible relationships. The number of sentences, number of words and sentences with citations correlation is just over 0.1 for all the citation metrics. Whilst this is weak it is worth considering other research (Falagas et al, 2013) where they state that the number of citations is related to article length and in that research they used Spearman correlation. Of the altmetrics, Mendeley shows the highest value, 0.14 with number of sentences and number of words. Whilst the correlation and thus the relationship is most likely negligible it is still interesting to note that the Altmetric score and Twitter have negative values compared to the others.

Splitting the data based on the Altmetrics score creating a high Altmetric score set, which included all data with a score of 50 or greater, resulted in correlation values of 1.5 to 2.2 between citations and repeated citations but none of these were significant at the 0.01 or 0.05 level. Looking at the low data set in Table 4, all values of Altmetric score less than or equal to 0.25, resulted in slightly stronger positive correlations being seen between Mendeley and number of sentences, total number of words and sentences with citations. The Altmetric score itself showed a stronger negative correlation with total

Table 4 Kendall Correlation Data set with Altmetric score <=0.25

	Scopus	CrossRef	PMC	Mendeley	Twitter	Altmetric Score
Number of Sentences	-	-	-	0.22	-	-0.18
Total number of Words	0.1	0.15	0.18	0.24	-0.13	-0.18
Sentences with Citations	0.19	0.16	0.18	0.19	(-0.06)	-0.1

Table 5 Kendall Correlation Full Data Set

	Repeated Citations	Sentences with Multiple Citations	Missing Citations
References	0.44	0.59	0.2
Sentences with multiple citations	-	-	0.27

number of words and number of sentences although these were still under 0.2. In general the correlations between the article level metrics themselves were slightly lower for this data set with the exception of Scopus and PMC which increased to 0.8 (not shown).

We examined content quantities for relationships that might be worth further exploration. A number of the relationships would be expected e.g. as the number of sentences goes up so would the total number of words in the article. One relationship that would not necessarily be expected is: As the number of references increases so does the number of repeated citations and sentences with multiple citations, see Table 5. As sentences with multiple citations increase there also seems to be a weak positive relationship with the number of references that are missing i.e. not cited in the main body.

6 Discussion and Conclusions

This research aims to understand relationships between traditional and new article level metrics and if relationships exist between these and content quantities. The literature review highlighted some of the potential problems with these traditional and newer metrics when they are used to indicate research quality. Our results did find that there is a relationship between citations and Mendeley readership counts. The relationship strength we found was comparable and in some cases lower than other studies but that could potentially be due to our smaller sample size and also our use of Kendall correlation as a measure and not Spearman correlation.

From the literature review it seems that the relationship between the newer and traditional metrics like citation is not clear and it cannot be said that one drives the other. We saw no relationship between citations and Twitter counts or Altmetric scores indicating that Almetric scores and Twitter are a measure of something else or there are other factors that may need to be accounted for. We found articles with high Altmetric scores or Twitter counts were what could be considered popular subject articles or about social media metrics themselves. We suggest a potential future area for investigation is in using topic analysis to take account of popular subjects as this may strengthen the understanding of relationships between traditional metrics, altmetrics and content quantity. One potential limitation with the analysis is that there was no time factor accounted for with the Altmetric score or Twitter counts. As more and more people use social web sites the higher usage may distort results when comparing across multiple years.

The relationship between content quantities and patterns is even less clear. With a full data set relationships seem negligible but within the lower Altmetric score data set patterns do emerge of relations with the number of sentences, length of article and how many sentences have citations. A limitation and an area for further exploration is that our data set became quite small (n=132) when looking at this low score. A larger data set at this lower end may yield stronger relationships. One relationship observed was that as the number of references increased so did the repeated citations, sentences with multiple citations and missing citations from the body. This could be a potential area of further investigation if more additional information can be gathered around citations.

Acknowledgements We thank Altmetric.com for the use of their API.

References

Altmetric (2015) Altmetric website. Available at:https://www.altmetric.com [Accessed: June2015]

Bornmann L (2014) Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. ArXiv e-prints 1407.8010, URL `http://arxiv.org/abs/1407.8010`

Costas R, Zahedi Z, Wouters P (2014) Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multi-disciplinary perspective. ArXiv e-prints 1401.4321, URL `http://arxiv.org/abs/1401.4321`

Coxhead A (2000) A new academic wordlist. TESOL Quarterly 34:213–238, DOI 10.2307/3587951

Donohue J, Erling E (2012) Investigating the relationship between the use of english for academic purposes and academic attainment. Journal of English for Academic Purposes 11(3):210–219, DOI 10.1016/j.jeap.2012.04.003

Falagas M, Zarkali A, Karageorgopoulos D, Bardakas V, Mavros M (2013) The impact of article length on the number of future citations: A bibliometric analysis of general medicine journals. PLoS ONE 8(2):e49,476, DOI 10.1371/journal.pone.0049476

Garfield E (1998) The impact factor and using it correctly. Der Unfallchirurg 101(6):413–414

Glänzel W, Debackere K, Thijs B, Schubert A (2006) A concise review on the role of author self-citations in information science, bibliometrics and science policy. Scientometrics 67(2):263–277, DOI 10.1007/s11192-006-0098-9

Gunther E (2011) Can tweets predict citations? Metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. Journal of Medical Internet Research 13(4):e123, DOI 10.2196/jmir.2012

Ha T, Tan S, Soo C (2006) The journal impact factor: Too much of an impact? Annals Academy of medicine Singapore 35:911–916

Haustein S, Costas R, V L (2015) Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. PLos One 10(3):e0120,495, DOI 10.1371/journal.pone.0120495

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proceedings of the National Academy of Science 102(46):16,569–16,572, DOI 10.1073/pnas.0507655102

Liu C, Xu Y, Wu H, Chen S, Guo J (2013) Correlation and interaction visualisation of altmetric indicators extracted from scholarly social network activities: dimensions and structure. Journal of Medical Internet Research 15(11):e259, DOI 10.2196/jmir.2707

Nomaler Z, Frenken K, Heimeriks G (2013) Do more distant collaborations have more citation impact? Journal of Informetrics 7(4):966–971, DOI 10.1016/j.joi.2013.10.001

PLoS (2015) PLoS website. URL `http://www.plos.org`, accessed: June2015

Plume A, Weijen D (2014) Publish or perish? The rise of the fractional author. Research Trends [online] URL http://www.researchtrends.com/issue-38-september-2014/publish-or-perish-the-rise-of-the-fractional-author/, accessed: September2015

Seglen P (1992) The skewness of science. Journal of American Society for Information Science 43(9):628–638, URL http://arxiv.org/abs/0912.4188

Vongpumivitch V, Huang J, Chang Y (2009) Frequency analysis of the words in the academic word list (AWL) and non-awl content words in applied linguistics research papers. English for Specific Purposes 28(1):33–41, DOI 10.1016/j.esp.2008.08.003

Ware M (2013) PRC peer review guide. Tech. rep., Publishing Research Consortium, URL http://publishingresearchconsortium.com/index.php/prc-guides-main-menu/155-peer-review-an-introduction-and-guide, accessed: September 2013

Ware M, Mabe M (2015) The STM report an overview of scientific and scholarly journal publishing[online]. Tech. rep., International Association of Scientific, Technical and Medical Publisher, URL http://digitalcommons.unl.edu/scholcom/9/, accessed: September 2015

Whitehouse G (2001) Citation rates and impact factors: should they matter? The British Journal of Radiology 74:1–3, DOI 10.1259/bjr.74.877.740001

Zahedi Z, Costas R, Wouters P (2014) How well developed are altmetrics? A cross-disciplinary analysis of the presence of alternative metrics in scientific publications. Scientometrics 101:1491–1513, DOI 10.1007/s11192-014-1264-0

Zhang G (2015) It is suggested that ... or it is better to .... ? Forms and meanings of subject it-extraposition in academic and popular writing. Journal of English for Academic Purposes 20:1–13, DOI 10.1016/j.jeap.2015.02.004