Scientometrics (2017) 110:243–251
DOI 10.1007/s11192-016-2161-5

The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts

Babak Sohrabi[1] · **Hamideh Iraj**[1]

Received: 3 July 2016 / Published online: 18 October 2016

Abstract This paper investigates an association between two new variables and citations in papers. These variables include the abstract ratio (the sum of repetition of keywords in abstract divided by abstract length) and the weight ratio (the frequency of paper's keyword per journal). The data consist of 5875 papers from 12 journals in education: three journals from each SCImago quartile. The researchers used semi-continuous regression to model the data and measure the impact of the proposed variables on citations. The results revealed that both abstract ratio and weight ratio are statistically significant predictors of citations in scientific articles in education.

Keywords Scientometrics · Bibliometrics · Citation analysis · Citation prediction

Introduction and background

Citation as a proxy, measures the quality of academic journals and researchers' impact. It is the building block of a journal's impact factor, researchers' h-index and several other measures. This measure may gain greater importance in the future. Bertsimas et al. (2015) paper, suggested moneyballing academia and using quantitative methods for key decisions in giving tenure and research grants, referred as "tenure analytics" (Bertsimas et al. 2015). These quantitative methods rely on citations so the academia is getting even deeper into citations.

The pressure for publishing highly cited papers went up in the recent years due to the competitiveness of research grants based on the researchers' impact. As a result, finding the variables that affect paper citations has been interesting for the publishers and the authors of scientific papers.

✉ Babak Sohrabi
bsohrabi@ut.ac.ir

[1] University of Tehran, Tehran, Islamic Republic of Iran

The effect of papers' bibliometrics on the citation has been studied partly because they are available at the acceptance of a paper and do not change over time. Although bibliometrics does not explain the whole story of citations nor does it claim to discover causal relationships (Robson and Mousques 2016), predicting citations in papers has become popular in academia.

Many researchers so far discovered bibliometric variables affecting citation counts. According to Yu et al., four factors affect paper citations: the authors, the journal, the research discipline and the papers themselves (Yu et al. 2014).

Author-related predictors include the number of articles and the number of citation for the first author, the number of articles and the number of citations for the last author, the number of institutions as well as the quality of the first author's institution (Fu and Aliferis 2010). Other predictors include gender (Robson and Mousques 2016) and standing of authors (Bornmann et al. 2012; Robson and Mousques 2016) and their country of residence (Robson and Mousques 2016).

Since researchers pay attention to core journals first and then they proceed toward papers in less reputable journals, in a positive loop highly cited journals receive more citations. In other words, journal's impact factor affects future citations to the journal papers. (Yu et al. 2014) In addition to that, the language of the journal is a predictor of citation (Bornmann et al. 2012).

The research discipline is yet another significant predictor of citations meaning that different disciplines have different citation patterns (Yu et al. 2014; Stegehuis et al. 2015; Robson and Mousques 2016; Brizan et al. 2016).

As mentioned before, the bibliometric variables include year of publication (Bornmann et al. 2014; Robson and Mousques 2016), number of pages (Bornmann et al. 2014; Robson and Mousques 2016), number of authors, number of references (Bornmann et al. 2014; Robson and Mousques 2016) or it is an open-source available (Robson and Mousques 2016).

Text-based predictors include paper abstract length (Robson and Mousques 2016), title length (Jacques and Sebire 2010; Habibzadeh and Yadollahie 2010; Paiva et al. 2012; Falahati Qadimi Fumani et al. 2015; Robson and Mousques 2016), combinatorial titles (separated by a hyphen or a colon) (Jacques and Sebire 2010; Rostami et al. 2014), having at least two keywords other than the words in the title (Rostami et al. 2014), the presence of an acronym in the title (Jacques and Sebire 2010), the number of punctuation marks in the title (Falahati Qadimi Fumani et al. 2015), whether time or place of the study was mentioned in the title (Jacques and Sebire 2010; Paiva et al. 2012; Alimoradi et al. 2016), whether the results were mentioned in the title or the method (Paiva et al. 2012)

The researchers explored two new variables in the realm of education journals: keyword repetitions in abstract and the frequency of keywords per journal. The goal of the current study is to figure out if there is any significant association between any of the two variables and citation counts. The findings help researchers and editors gain an insight into the intelligent use of keywords in scientific papers to increase citations.

Research methodology

Research design

The current research uses a quantitative research method. Further, the impact of two proposed variables on citations is measured through statistical techniques.

Data collection

The data were collected from the Scopus database for the years 2000–2015 in May 2016. The researchers considered SCImago rankings for education journals (SCImago 2015) and selected three journals of each quartile randomly. These 12 journals and their ISSNs include:

- *Quartile 1* Computers and Education (03601315), Internet and Higher Education (10967516) and Teaching and Teacher Education (0742051X)
- *Quartile 2* Asia Pacific Education Review (15981037), Teacher Development (17475120) and Teaching in Higher Education (14701294)
- *Quartile 3* Perspectives in Education (02582236), History of Education (14645130) and Journal of Continuing Higher Education (19484801)
- *Quartile 4* Computers in Education Journal (10693769), Evaluation and Research in Education (09500790) and International Journal of Pedagogies and Learning (18334105)

Only original research articles (type='article') were considered for the study because earlier works found "the type of study" a predictor of citation. Therefore, the researchers controlled for the type of study by fixing the value to type='article' (Fu and Aliferis 2010; Stegehuis et al. 2015). The dataset had 5875 usable papers. Papers with incomplete bibliometrics were excluded from the dataset such as papers without abstract/reference/keywords.

Variables calculation

The following sections briefly elaborate the research themes i.e. abstract ratio and weight ratio calculations.

Abstract ratio

The abstract ratio is the sum of keyword repetitions in the paper abstract divided by the length of abstract. The researchers calculated the abstract ratio using the following steps: First the variable "author.keyword" was split into single keywords and created them as k1, k2…Then, for each keyword, a new variable was created named k1_abstract, k2 abstract…These variables store the count of each keyword repetition in abstract. Then, the sum of these counts is stored in a variable called abstract_sum, which was further divided by the length of the abstract to create the variable "abstract_ratio". This variable was an input in the data analysis phase.

Weight ratio

The Weight ratio is the sum of the frequency of each keyword in a specific journal. It was calculated for each paper in the following order: First, the researchers used k1, k2… from the previous section. Second, a separate aggregate table at the journal level was built showing the relative frequency of each keyword in a journal. Third, an extra variable for each keyword named k1_weight, k2_weight… retrieved the relative frequency for each paper-keyword from the aggregate table. Fourth the variable weight_ratio was calculated as the sum of weight variables.

The dataset now looks like the schema in Table 1.

Table 1 Dataset variables

Paper title	Author keywords	K1	K1_abstract	K1_weight	K2	K2_abstract	K2_weight	…	Abstract_ratio	Weight_ratio
	K1; k2; k3; …	Blended learning		0.1						
					Blended learning		0.1			

Table 2 Descriptive statistics of predictors

	No_authors	Article_age	Page.count	No_references	Abstract_length	Title_length	Abstract_ratio	Numkeys	Quartile	Weight_ratio
Min	1.000	1.000	1.00	1.00	27.0	3.00	0.000000	1.000	q1:3978	0.0004717
1st Quartile	1.000	3.000	10.00	30.00	115.0	10.00	0.006173	4.000	q2:1163	0.0022075
Median	2.000	5.000	12.00	42.00	151.0	12.00	0.023622	5.000	q3:599	0.0060893
Mean	2.399	5.756	12.52	46.53	158.5	12.73	0.029025	4.661	q4:135	0.0217250
3rd Quartile	3.000	8.000	15.00	58.00	189.0	15.00	0.044444	5.000	NA	0.0256410
Max	13.000	16.000	36.00	306.00	582.0	35.00	0.201923	14.000	NA	0.2099057

Table 3 Citation prediction results

Target Transformed target	Model 1 (logistic regression) cited.by.positive cited.by.positive Beta coefficient	*P* value	Model 2 (linear regression) cited.by log(cited.by) Beta coefficient	*P* value
Abstract_length	0.002	<0.05	0.002	<0.05
Page.count	−0.052	<0.05	−0.008	~0.05
Title_length	−0.007	>0.05	−0.012	<0.05
Abstract_ratio	5.216	<0.05	1.512	<0.05
No_references	0.014	<0.05	0.004	<0.05
No_authors	0.054	>0.05	0.029	<0.05
Numkeys	−0.001	>0.05	−0.009	>0.05
Article_age	0.651	<0.05	0.166	<0.05
Weight_ratio	3.58	<0.05	2.991	<0.05
Quartileq2	−1.173	<0.05	−0.947	<0.05
Quartileq3	−2.238	<0.05	−1.573	<0.05
Quartileq4	−2.55	<0.05	−1.776	<0.05

Data analysis

Data analysis was done using R version 3.2.0 (2015-04-16) and "car" package (Fox and Weisberg 2011). The inputs for data analysis were 10 variables extracted from the literature review including abstract length (in words), page count, title length (in words), number of references, number of authors, number of keywords, article age (in year), the SCImago quartile (q1, q2, q3 or q4), abstract ratio and weight ratio.

The logistic regression (Model 1) and the least-squares linear regression (Model 2) (Fox and Weisberg 2011) were used in the current study to predict the target variable citation. It is a positive discrete variable. For the current study, it was treated as a continuous variable since it has a large number of distinct values. These two models were preferred because of being simple and interpretable. In addition, they allow a mix of numerical and categorical variables as predictors and hence; they were widely used in earlier studies (Thelwall and Wilson 2014).

Results

Descriptive statistics of model predictors are depicted in Table 2.

As seen, the researchers conducted a semi-continuous prediction. The dataset was split into two parts: zero citations and positive citations. A logistic regression predicted whether the paper will be cited or not (zero citation versus positive citation) and a linear regression predicted citations among those cited more than zero. Furthermore, the use of the semi-continuous model allowed the researchers to consider skewness of the target variable (citation) and the high number of zero citations.

Model 1, as shown in Table 3, predicted if the paper will be cited or not (AIC = 3174.7). The researchers coded zero citations as zero and positive citations as one

in the binary variable cited.by.positive. According to Model 1 results, both the abstract ratio [p value = 0.00416, beta = 5.25, confidence interval = (1.66, 8.83)] and the weight ratio are significant [p value = 0.03429, beta = 3.57, confidence interval = (0.34, 6.94)].

Model 2 predicted citation in papers of positive citations. (R-squared = 0.4, Adjusted R-squared = 0.4, model p value = 2.2e−16) Here, the researchers used box-cox transformation to find the right lambda and meet the assumptions of the linear model (Fox and Weisberg 2011, pp. 303–309).

As depicted in the Model 2 results, both the abstract ratio [p value = 0.009267, beta = 1.51, confidence interval = (0.37, 2.65)] and the weight ratio [p value = 7.12e−11, beta = 2.99, confidence interval = (2.09, 3.89)] are significant. Considering the logarithm of the citation variable in Model 2 is consistent with previous researchers' recommendations such as (Thelwall and Wilson 2014). The results of the semi-continuous method revealed that the abstract weight ratios are significant variables in predicting if the paper will be cited (Model 1) and how many citations it will receive (Model 2).

Answering the research questions, the abstract ratio and the weight ratio are statistically significant in the two models. They have small p values, greater than one odds ratio in Model 1, positive coefficient in Model 2.

Challenges

The first major challenge the current work came across was the poor quality of the Scopus dataset. The researchers spent a lot of time on journals' selection and data cleaning. There were cases where the journal bibliometrics was useless meaning journals in the dataset quite often were devoid of abstracts or references or keywords. Other problems existed as well: There were instances of papers with the wrong start or end pages (leading to a negative page number). The reference variable format was inconsistent too; in some journals, a semicolon separates the references while in the rest, a comma is the separator. The data quality problem prevented the researchers from using the data on a larger scale.

The feature engineering and interpreting the results were other challenges in the course of the study. There were several bibliometric predictors found during the literature review but there were not a priori knowledge which variables work out in the context of this study because the bibliometric predictors are conceptually disconnected from citation counts. As such, making the sense of data and interpreting the results was difficult.

Discussion and conclusion

The current study investigated the effect of two variables in predicting citation counts: keyword repetition in abstract and keyword frequency per journal. These two variables were significant predictors with positive coefficients. In other words, controlling for other variables, an increase per unit of these variables is associated with higher citation for a paper.

Among other variables, abstract length, page count, the number of references, the article age were significant in the two models. Category levels q2, q3 and q4 were also statistically significant. That makes sense because journals quartiles come from impact factors and journals impact factors come from the sum of citation counts for each journal. Their

negative coefficients and an increase in absolute value from q2 to q4 indicates that compared to q1 as the base level, citation counts decrease (less than one odds ratio) when we go from q1 to q2, q1 to q3 and q1 to q4 respectively.

The use of the semi-continuous regression allowed the researchers to investigate the effect of the two new variables from two different perspectives: whether they will be cited or not and how many times they will be cited. A variable may predict the former or the latter or any combination of the two target variables.

The results of the current study have the following implications: first, the abstract ratio is a significant predictor of citation count i.e. researchers can boost citations by repeated keywords in the abstract. This is important technically and conceptually. From the technical point of view, a phrase repetition in an abstract increases the chance of retrieval in a search engine (Ale Ebrahim et al. 2013). Conceptually, when keywords are relevant to an abstract, increases the chances of its usefulness to readers.

Second, the significance of the weight ratio means that papers with frequent keywords at journal level are more likely to be cited. For example, if a journal's frequent keywords are A, B and C, papers with these keywords are cited more than those with less frequent keywords. This makes sense. When a journal is reputable in one topic or keyword, researchers are more likely to cite it in their papers. Therefore, citations go up by publishing in relevant journals. Although this is not a new suggestion, it was untested quantitatively before the current research.

The current study contributes to the body of knowledge by revealing the importance of keywords in citation counts in academic journals. The researchers found that papers with repeated keywords in abstract and frequent keywords at the journal level are more likely to be cited, a fact authors and editors neglected so far. In a nutshell, authors can boost the citation by carefully written abstract and keywords and editors can help make abstract and keywords as relevant as possible.

References

Ale Ebrahim, N., Salehi, H., Embi, M., Habibi Tanha, F., Gholizadeh, H., Motahar, S., et al. (2013). Effective strategies for increasing citation frequency. *International Education Studies, 6*(11), 93–99.

Alimoradi, F., Javadi, M., Mohammadpoorasl, A., Moulodi, F., & Hajizadeh, M. (2016). The effect of key characteristics of the title and morphological features of published articles on their citation rates. *Annals of Library and Information Studies, 63*, 74–77.

Bertsimas, D., Brynjolfsson, E., Reichman, S., & Silberholz, J. (2015). OR Forum—Tenure analytics: Models for predicting research impact. *Operations Research, 63*(6), 1246–1261.

Bornmann, L., Leydesdorff, L., & Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics, 8*, 175–180.

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics, 6*, 11–18.

Brizan, D., Gallagher, K., Jahangir, A., & Brown, T. (2016). Predicting citation patterns: defining and determining influence. *Scientometrics, 108*, 183–200.

Falahati Qadimi Fumani, M., Goltaji, M., & Parto, P. (2015). The impact of title length and punctuation marks on article citations. *Annals of Library and Information Studies, 62*, 126–132.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks CA: Sage. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*, 257–270.

Habibzadeh, F., & Yadollahie, M. (2010). Are shorter article titles more attractive for citations? Cross-sectional study of 22 scientific journals. *Croatian Medical Journal, 51*(2), 165–170.

Jacques, T., & Sebire, N. (2010). The impact of article titles on citation hits: an analysis of general and specialist medical journals. *Journal of the Royal Society of Medicine Short Reports, 1*(1), 1–5.

Paiva, C., Lima, J., & Paiva, B. (2012). Articles with short titles describing the results are cited more often. *Clinics, 67*(5), 509–513.

Robson, B., & Mousques, A. (2016). Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts. *Environmental Modelling and Software, 75*, 94–104.

Rostami, F., Mohammadpoorasl, A., & Hajizadeh, M. (2014). The effect of characteristics of title on citation rates of articles. *Scientometrics, 98*(3), 2007–2010.

SCImago. (2015). *Journal Rankings on Education - Scimago Journal & Country Rank*. Retrieved from http://www.scimagojr.com/journalrank.php?category=3304.

Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications. *Journal of Informetrics, 9*, 642–657.

Thelwall, M., & Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *Journal of Informetrics, 8*, 963–971.

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics, 101*, 1233–1252.