The correlation between content novelty and scientific impact

Shiyun Wang
School of Information Management
Wuhan University
Wuhan, Hubei, China
wangsy2@whu.edu.cn

Jin Mao†
School of Information Management
Wuhan University
Wuhan, Hubei, China
maojin@whu.edu.cn

Yaxue Ma
School of Information Management
Nanjing University
Nanjing, Jiangsu, China
myx_vicky@163.com

ABSTRACT

Novel research drives scientific breakthroughs but also has higher uncertainty of being recognized by citation count based metrics. This study proposed two indicators to measure the content novelty of a paper based on the knowledge entities it contains, and explored the relationship between content novelty and scientific impact of papers. It is found that content novelty is negatively correlated with citation impact in our dataset. Our findings suggest that science policy in favor of citation count based impact may be biased against novel research.

CCS CONCEPTS

•Applied computing~Document management and text processing

KEYWORDS

Novelty, Scientific impact, Research evaluation, Bibliometrics

1 Introduction

Novelty has long been considered as one driver of scientific breakthrough and of economic growth [1, 2]. Nonetheless, researches with high novelty also face higher risk [2]. They are often hard to be accepted by peer reviews in a short period of time, resulting in a higher probability of being published in low-impact journals. The citation count of papers with high novelty may be very few, and it may take a longer time to receive a major impact. While one major factor of research evaluation in contemporary science is citation count, which might underestimate the value of these studies in the early years.

A few studies have proposed some indicators to measure the novelty of papers. For example, Uzzi et al. [3] calculated the atypicality of journal pairs in reference list of a paper. However, these indicators use the journals of references to represent knowledge component of an article, rather than the knowledge content of the article.

The present paper offers new measures of novelty by exploiting knowledge content of papers. Innovation is often not a flash of light, but a result of standing on the shoulders of giants. Building on this idea, we define research that draws on new knowledge content that compared to its references as novel, and develop two indicators to measure the content novelty of a paper. Applying the new indicators of novelty, we further explore the correlation between novelty and citation impact.

2 Methods

2.1 Dataset

The data was collected from an open-access PubMed dataset[1]. Our analysis was performed on 634,738 journal articles published in 2009 with at least one reference. The papers were categorized into six domains based on the Science-Metrix classification scheme[2], including Applied Sciences, Arts & Humanities, Economic & Social Sciences, General, Health Sciences and Natural Sciences.

2.2 Content novelty indicators

The knowledge content of a paper is represented by the Pubtator Central[3] entities and the pairwise combination of entities in the paper. The PubTator Central system provided biomedical concepts such as genes, chemicals that were automatically extracted from each PubMed abstract. The F1 score of this system is higher than 80% [4]. We obtained entities of each article in our dataset by searching PMID in the system via API.

We determined the novelty degree of a paper as the proportion of its new knowledge entities and new knowledge entities pairs that were not appeared in its references. We compared the entities in two sources by exactly matching, which means the same entities in the two sources should be exactly identical. Formally, the two indicators were computed as follows:

(1) The proportion of new knowledge entities in a paper p:

$$ent_p = \frac{n_{pi}}{n_p} \tag{1}$$

† Corresponding author.

[1] https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[2] https://science-metrix.com/?q=en/classification
[3] https://www.ncbi.nlm.nih.gov/research/pubtator/

The n_p is the number of knowledge entities in paper p, while n_{pi} is the number of new knowledge entities in paper p that were not occurred in its references.

(2) The proportion of new pairwise combination of knowledge entities in paper p:

$$com_p = \frac{r_{pi}}{r_p} \tag{2}$$

The r_p is the number of distinct pairwise combination of knowledge entities in paper p, while r_{pi} is the number of new pairwise combination of knowledge entities in paper p that were not appeared in its references.

2.3 Scientific impact

Citation count has often been applied to evaluate the scientific impact of publications. In this paper, we also explored the relationship between content novelty and citation counts of papers. We ranked the novelty values of all papers in ascending order. For each decile of novelty, we calculated the average number of citations each paper received over a 3-year period after the publication year, and the proportion of top10% highly cited papers. These processes also applied for each domain of our dataset separately to observe the differences among domains. In addition, we used Pearson correlation coefficient to measure the strength of association between content novelty and scientific impact. We considered short-term impact (1-year citations), medium-term impact (3-year citations and 5-year citations) and long-term impact (10-year citations) of papers.

3 Results

3.1 Scientific impact of different novelty groups

Figure 1 presents the mean of citation counts for each decile of novelty, where novelty was measured by *ent* and *com* indicators respectively. We can observe that citations decrease significantly as the rise of novelty. The paper in the last decile of novelty has more than 10 citations less than the paper in the first decile on average, either in *ent* or *com* measured novelty. A slight increase of citation counts is observed in the 10-20 percentile group of novelty that measured by *ent*. These patterns are robust across the major domains in our dataset.

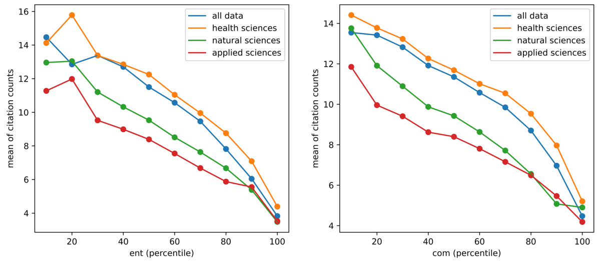

Figure 1: Average citation counts of papers in different novelty groups. We only provided the results of three main domains in our dataset.

The proportion of top10% highly cited papers declines with the deciles of novelty, as shown in Figure 2. The result is also robust across domains in our dataset. Only Health Sciences exhibits a slightly different pattern that the proportion of top10% papers has increased in the 10-20 percentile group of *ent* measured novelty.

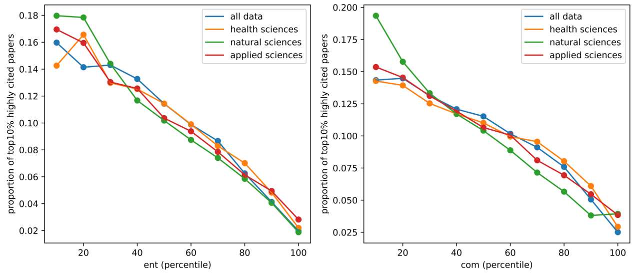

Figure 2: Proportion of top10% highly cited papers in different novelty groups.

3.2 Relationship between content novelty and scientific impact

Table 1 presents the Pearson correlation coefficients between content novelty indicators and different impact of papers. It shows that content novelty of papers has significantly negative correlation with scientific impact. However, the correlation coefficients are not very large, ranging from 0.078 to 0.130.

TABLE 1. The Pearson correlation coefficient between content novelty and scientific impact.

Content novelty	**Citation impact**			
	1-year	3-year	5-year	10-year
ent	-0.114	-0.130	-0.126	-0.100
com	-0.087	-0.101	-0.098	-0.078

Note: The p-values were all smaller than 0.001.

4 Discussion and Conclusions

In this article, we have introduced two indicators to measure the content novelty of papers, which were operationalized as the proportion of new entities and new pairwise combination of knowledge entities in the paper that were not appeared in its references. We find that the content novelty is negatively correlated with the citation impact, which indicates that the widely used citation count based measures are biased against novel research, and thus may fail to recognize novel research in science [2]. More indicators combining knowledge entities and pairwise combination of entities could be proposed to comprehensively measure the content novelty of papers and to evaluate research in science.

ACKNOWLEDGMENTS

This study is supported by NSFC grants, No. 71921002 and No. 71804135.

REFERENCES

[1] Fontana, M., Iori, M., Montobbio, F., and Sinatra, R. 2020. New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy* 49(7), 104063. doi: 10.1016/j.respol.2020.104063.

[2] Wang, J., Veugelers, R., and Stephan, P. 2017. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy* 46(8), 1416–1436. doi: 10.1016/j.respol.2017.06.006.

[3] Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. 2013. Atypical Combinations and Scientific Impact. *Science* 342(6157), 468 – 472. doi: 10.1126/science.1240474.

[4] Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* 47(W1), W587–W593. doi: 10.1093/nar/gkz389.