DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.3, AUG. 2021

Understanding more about non–patent references

Xiaojun Hu[a,b]* , Yushuang Lv[a], Zixuan Zhao[a], Xian Li[a]

a. Medical Information Center, Zhejiang University School of Medicine, Hangzhou, China

b. Department of Neurology of Affiliated Hospital 2, Zhejiang University School of Medicine, Hangzhou, China

ABSTRACT

Non-patent references (NPRs) play an important role in the interaction between science and technology. In this contribution, we compare the characteristics of NPRs traced from patents on "metabolic engineering" to those received by the 10% most-cited articles (cited in scientific publications and excluding NPRs) in the same field. We observe that there is a low rate of co-citing articles between the two groups, and many NPRs have a better citation performance than TOP 10% articles. Our results provide preliminary clues to discover new characteristics of NPRs and to better understand the interface between science and technology.

KEYWORDS

Citation studies; Non-patent references; Metabolic engineering; Co-citing articles; Sci-tech boundary; Patent data analysis

1 Introduction

Non-patent references (NPRs) articles can be used as proxies to trace knowledge transition between science and technology (Hu & Rousseau, 2018). For this reason, NPRs have received more and more attention in recent years (Callaert et al., 2012; Liaw et al., 2014; Sung et al., 2015). The underlying idea of these investigations is that the more characteristics of NPRs are uncovered, the better our understanding of the boundary and the mutual influence between science and technology.

It is known that the percentage of NPRs in the interface between science and technology is very low. Tijssen (2010), for instance, found that there were only 31 WoS Subject Categories with'non-patent references' (NPR) scores above 0.25%, while van Raan (2017) found that only 3-4% WoS publications can be identified as NPRs. Hence, one wonders if the group of NPRs has atypical characteristics vs the main group of articles in the WoS, or even vs. the group of most cited articles in science.

The purpose of this contribution is to study and try to find answers to the following questions:

(1) Are NPRs less or more cited in scientific publications than non-NPR TOP 10% most cited articles on a certain topic?

(2) What are the differences between NPRs and non-NPR TOP 10% articles?

*Corresponding author: xjhu@zju.edu.cn

2 Definition of the terms NPRs and TOP 10% articles as used in this investigation

In order to find characteristics which stay usually below the surface, a large set of citation data, containing more than 114,000 records from the WoS were downloaded and analyzed.

An NPR-article: In this investigation we define an NPR as a scientific article that was cited by more than one patent. To focus on the characteristics of active items in the NPR group, we use "the first five years (not including the publication year)" as a citation window to refine NPRs included in this study.

A TOP 10% article in scientific communication (TOP 10%): A scientific article that belongs to the TOP 10% most-cited articles of its publication year, and is not an NPR. Hence it received?no more than one patent citation during the first five years after its publication.

3 Methodology

(1) Research field. We focus on metabolic engineering, it is an emerging field dealing with the practice of optimizing genetic and regulatory processes within cells to increase these cells' production of a certain substance. This field contains a huge amount of topics with potentially patentable innovations (Yang et al., 1998)

(2) Data source and collection. All data were extracted from three databases:Web of Science, Derwent Innovation Index, and Derwent Innovation, using the search strategy: TS= "metabolic engineering" AND DT= "article" , and that starting from the year 2001.

(3) Removing the overlapping items in two sets. We identify overlapping items in the NPR set and the TOP 10% set, and remove them.

(4) Citation data. For all articles in the NPR and the TOP 10% sets, we determined the number of received citations as available in the WoS. The two groups together received a total of 114,332 citations, all of them were downloaded to be analyzed.

(5) Yearly rate of citation diffusion. The yearly rate citation diffusion defined as the number of received citations from outside the ESI-field to which an article belongs divided by the total number of received citations, calculated per year.

(6) Time series analyses. We apply diachronous and synchronous time series analyses to study the articles in the NPR and TOP 10% sets. The yearly number of citations, the yearly rates of citation diffusion outside ESI-fields based on cumulative data were determined. In these time series, Y is the publication year of an article, years Y+1, Y+2, Y+3,...,Y+n are the following years, and Y-1, Y-2, ... are preceding years. Particularly, in the synchronous time series, we considered all articles in the NPR set and the TOP 10% set published during a period [Y-x, Y].

4 Results

We find considerable differences between articles in the two sets.

4.1 The differences of yearly number of received citations between NPRs and TOP 10% articles

Figures 1 to 2 show time series for the NPR and TOP 10% groups. These time series are: the yearly mean and median number of received citations. Clearly, the average values shown in Figure 1 of the yearly number of received citations of articles in the NPR group are much higher than those of the TOP 10% group. Yet, this difference disappears when considering median values. Moreover, one can see that for the average citation values for NPRs, the curve shows a steep increase, followed by a small decrease, and again a small increase (although one expects a decrease as articles are already more than ten years old). In contrast, the trend of average yearly citations for the curve of TOP 10% is decreasing over time.

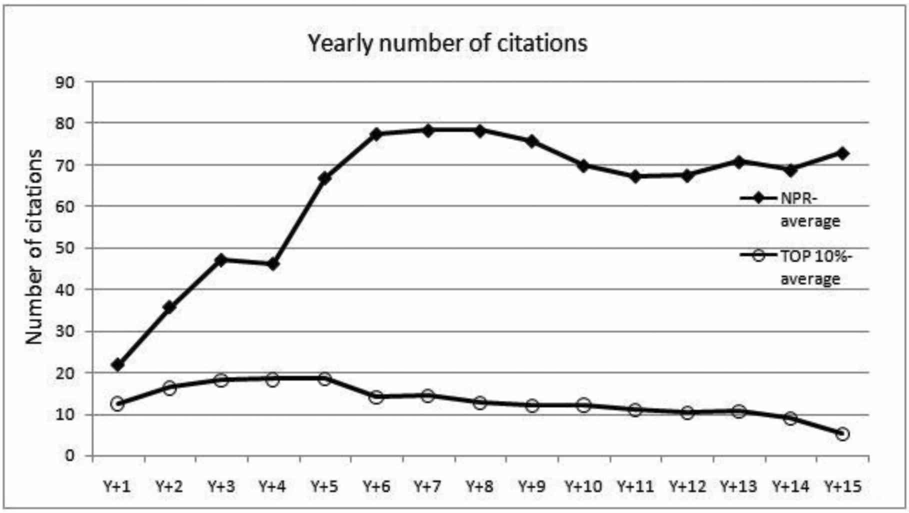

Figure 1 Trend lines of the mean number of citations for NPRs and TOP 10% articles

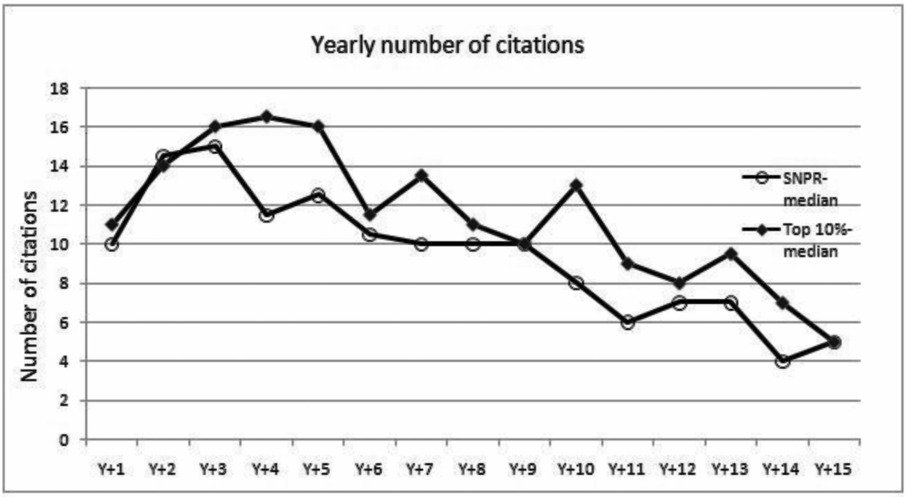

Figure 2 Trend lines of the median number of citations for NPRs and TOP 10% articles

4.2 The yearly rates of citation diffusion to outside ESI–fields, based on cumulative citations

As shown in Figure 3, the yearly rates of citation diffusion outside ESI-fields of NPRs are higher than the corresponding yearly values of TOP 10% articles. Furthermore, the cumulative yearly rates of citation diffusion outside ESI-fields of both NPRs and TOP 10% articles are growing over time, with trend lines y=0.0022x+0.6068, and y=0.0115x+0.4346, respectively. Here y denotes the diffusion rate and x denotes time since publication.

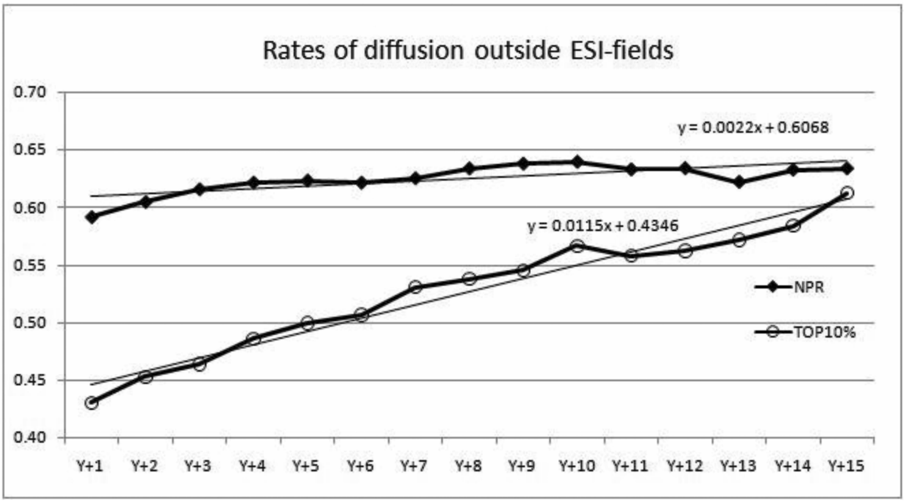

Figure 3 The mean values of yearly rates of citation diffusion outside ESI-fields for the two groups of articles (cumulative)

4.3 Citation performance of the articles in the two groups in fixed years based on synchronous time series

Tables 1- 4 display the mean and median number of citations in the fixed years Y+5, Y+10, Y+15 for NPR and TOP 10% articles in different publication periods. These tables provide additional information illustrating the events shown in Figures 1 to 3.

For the periods [2011 - 2012], [2006 - 2010] and [2001 - 2005], the citation values in fixed years of articles in the NPRs group are not always higher than those of the corresponding values of the TOP 10% group, especially for the median values. As shown in Tables 1-2, for articles published in the period [2001; 2005], the cumulative number of citations in the years Y+5, Y+10 of the NPRs are lower than those for the TOP 10%, and this for mean as well as for median values. Yet, the performance of NPRs is better in the year Y+15. However, most NPRs published in the oldest period, namely [1978 - 2000] received a much higher number of citations in the years Y+5, Y+10, and Y+15 compared to all other cases, which suggests that the older NPRs, play the most important role in the results shown in Figures 1-2.

Table 1 Number of citations in three fixed years after publication of the two groups based on synchronous time series (cumulative mean)

Group	Citation year Publication Period	Y+5	Y+10	Y+15
NPR	[2011 – 2012]	220		
	[2006 – 2010]	91.5	126.5	
	[2001 – 2005]	64	129.4	266.2
	[1978 – 2000]	277.91	778.36	1182.47
TOP 10%	[2011 – 2012]	102.19		
	[2006 – 2010]	70.93	153.43	
	[2001 – 2005]	81.67	145.67	170.11

Table 2 Number of citations in three fixed years after publication of the two groups based on synchronous time series (cumulative median)

Group	Citation year Publication Period	Y+5	Y+10	Y+15
NPR	[2011 – 2012]	220		
	[2006 – 2010]	40.5	74.5	
	[2001 – 2005]	29	74.5	381
	[1978 – 2000]	90	133	168
TOP 10%	[2011 – 2012]	82		
	[2006 – 2010]	68	137.5	
	[2001 – 2005]	70.5	136	168

Table 3 Yearly rates of citation diffusion outside ESI-fields in three fixed years after publication for the two groups based on synchronous time series (cumulative mean)

Group	Citation year Publication Period	Y+5	Y+10	Y+15
NPR	[2011 – 2012]	0.268		
	[2006 – 2010]	0.738	0.669	
	[2001 – 2005]	0.415	0.480	0.450
	[1978 – 2000]	0.642	0.664	0.650
TOP 10%	[2011 – 2012]	0.49		
	[2006 – 2010]	0.52	0.59	
	[2001 – 2005]	0.47	0.54	0.61

Table 4 Yearly rates of citation diffusion outside ESI-fields in three fixed years after publication for the two groups based on synchronous time series (cumulative median)

Group	Citation year Publication Period	Y+5	Y+10	Y+15
NPR	[2011 – 2012]	0.268		
	[2006 – 2010]	0.845	0.718	
	[2001 – 2005]	0.393	0.485	0.389
	[1978 – 2000]	0.664	0.690	0.660
TOP 10%	[2011 – 2012]	0.42		
	[2006 – 2010]	0.39	0.51	
	[2001 – 2005]	0.43	0.52	0.60

5 Discussions and conclusions

To compare the two special article "types" in a scientific context, we used articles originating from patents on the topic "metabolic engineering" starting from the year 2001 as NPRs, and TOP 10% articles of the topic "metabolic engineering" during the same period. Through a series of steps leading to a large citation database of more than 114,000 records, combining three databases WoS, DII, and DI, we present a novel framework to study the difference between TOP 10% most-cited articles in scientific realm and non-patent references traced from patent context, and their role in the evolution of scientific citations.

5.1 Citation performance of the NPR group deserves more attention

From Figures 1 to 3 we understand that the yearly number of received citations, the cumulative rates of citation diffusion outside ESI fields, suggested that the distinctive citation performance of NPRs may not be visible in the beginning period.

Another important observation is that the trends of yearly citation curves of NPRs and TOP 10% shown in Figure 1 indicate that in most cases, the citation life of NPRs is longer than that of TOP 10% articles; Yet, this difference disappears when considering median values (as shown in Figure 2), suggested that there are some outliers in the NPR set with a high and increasing number of citations (much higher than for articles in the TOP 10% group).

5.2 Older articles are major actors in the story about NPRs

From the synchronous time series shown in Tables 1 and 2, one may understand that in the group of NPRs, articles published in the oldest period [1978 - 2000], a share of 71% articles, plays an important role to produce the results that are shown in Figures 1-2. These results are indeed a reflection of the mean values (also median ones) in the fixed years Y+5, Y+10, and Y+15, which suggests that many older articles listed as NPRs belong to popular scientific work. Hence, they should receive special attention in investigations on NPRs. This observation, however, is in contrast to earlier findings that in emerging and developing fields the time lag is mostly relatively short (van Raan, 2017).

As a preliminary study towards new insights, we designed a methodology to process a series of data related to the science-technology border. We hope that our work will refresh traditional approaches.

One limitation of this study is that we focus on active NPRs and that, as this is just a case study, the observed characteristics need confirmation for other fields.

In conclusion, we have found answers to the research questions in the beginning. Our results provide preliminary clues to uncover more properties of NPRs. These articles can be considered a special group worthy of further informetric studies, leading to a better understanding of the boundary between science and technology.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, Grant Nos 71974167&71573225.The authors thank Ronald Rousseau for his help during the preparation of this article.

References

Callaert, J., Grouwels, J., &Van Looy, B. (2012) . Delineating the scientific footprint in technology: Identifying scientific publications within non–patent references. *Scientometrics*, *91* (2) , 383–398.

Hu, X. J., & Rousseau, R. (2018) . A new approach to explore the knowledge transition path in the evolution of Science & Technology: From the biology of restriction enzymes to their application in biotechnology. *Journal of Informetrics, 12* (3) , 842–857.

Liaw, Y. C., Chan, T. Y., Fan, C. Y., & Chiang, C. H. (2014) . Can the technological impact of academic journals be evaluated? The practice of non–patent reference (NPR) analysis. *Scientometrics*, *101* (1) , 17–37.

Sung, H. Y., Wang, C. H., Huang, M. H., & Chen, D. Z. (2015) . Measuring science–based science linkage and nonscience–based linkage of patents through non–patent references. *Journal of Informetrics*, *9* (3) , 488–498.

Tijssen, R.J.W. (2010) . Discarding the ʻBasic Science/Applied Science´ dichotomy: A knowledge utilization triangle classification system of research journals. *Journal of the American Society for Information Science and Technology*, *61* (9) , 1842–1852.

van Raan, A.F.J. (2017) . Patent citations analysis and its value in research evaluation: A review and a new approach to map technology–relevant research. *Journal of Data and Information Science*, *2* (2) , 13–50.

Yang, Y.T., Bennet, G.N., & San, K.Y. (1998) . Genetic and metabolic engineering. *Electronic Journal of Biotechnology*, *1* (3) , 134–141.