Comparing patent front-page and in-text references to science

Jian Wang[*] and Suzan Verberne [**]

[*] *j.wang@sbb.leidenuniv.nl*
0000-0003-0520-737X
Leiden Institute of Advanced Computer Science, Leiden University, Netherlands

[**] *s.verberne@liacs.leidenuniv.nl*
0000-0002-9609-9505
Leiden Institute of Advanced Computer Science, Leiden University, Netherlands

Patent citations to science provide a paper trail of knowledge flow from science to innovation, and have attracted a lot of attention in recent years. However, most studies rely on patent front-page references. We compare these two types of references and test whether they lead to the same analytical results regarding the relationship between science and innovation. Using a dataset of 33,337 USPTO biotech utility patents and their 860,879 in-text references and 637,570 front-page references to Web of Science journal articles, we found a remarkable low overlap between these two types of references. In-text references are more basic and have more scientific citations than front-page references. In-text references are less interdisciplinary but more novel than front-page references, but the differences are small. Furthermore, they lead to very different results when investigating how basicness, interdisciplinarity, novelty, and scientific citations affect patent citations.

1. Introduction

Since the pioneer work of Nunn and Oppenheim (1980), Narin and Noma (1985), and Tamada et al. (2006), sNPRs have been widely used for studying the interaction between science and technology (Hicks et al., 2000; Ke, 2020a, 2020b; Poege et al., 2019; Popp, 2017; Veugelers & Wang, 2019). However, most studies use patent front-page references and ignore patent in-text references. Front-page references are the references listed on the front page of the patent document, which are deemed as relevant prior art for assessing patentability. In-text references are references embedded in patent full text, serving a similar role as references in scientific papers.

Some recent studies have shown that in-text references embody information that is different from front-page references and provide a better proxy of knowledge flow (Bryan et al., 2020; Marx & Fuegi, 2020; Verberne et al., 2019; Voskuil & Verberne, 2021). Front-page references are generated because of patent applicants' legal duty to disclose prior art that is relevant for assessing the patentability of the focal invention. In comparison, in-text references are more like references in academic papers and may capture prior research that has enabled or influenced the focal invention but does not directly relate to patentability. Consistent with the process through which references are generated, several earlier studies based on surveys or interviews have suggested that front-page references may not represent a direct link between the citing patent and the cited scientific paper, but the cited scientific paper plays a more indirect role as a source of relevant background information (Callaert et al., 2014; Meyer, 2000; Nagaoka & Yamauchi, 2015; Tijssen et al., 2000). In their work that pioneered the analysis of patent front-page references, Narin and Noma (1985) stated that patent in-text references might be a better instrument for tracing knowledge flow from science to technology. Bryan et al. (2020) provided further evidence supporting this statement: First, using paper-patent-pairs (which consists of a scientific paper and a patent about the same biotech research output), references in the paired paper are much more likely to be cited by the paired patent in text rather than on front page. Second, using firm survey data, the number of a company's patent in-text references to science is more strongly correlated with its reliance on science as reported by its R&D manager. Furthermore, Marx and Fuegi (2020)

found that in-text references are older, less localized, less self-cited, more interdisciplinary, and more cited by future patents, compared with front-page references.

Taken together, prior studies suggest that in-text reference is a better proxy of knowledge flow and that studies using in-text reference may discover patterns different from studies using front-page references. Therefore, this paper will compare the difference between patent in-text and front-page references, and test whether results concerning science technology linkages will be different depending on which types of references are being used.

2. Data and method

2.1. Data

Our paper-patent-links data come from Voskuil and Verberne (2021). They trained the state-of-the-art BERT-based machine-learning models for extracting patent in-text references to scientific papers. The pre-trained BERT models were fine-tuned on a set of 1,952 hand-labelled references in 22 patent documents. The algorithm automatically classified words into three categories: (B) the beginning of a reference, (I) inside a reference, and (O) outside a reference. The accuracy of the algorithm is reflected in its prediction power for B and I labels (Ramshaw & Marcus, 1999; Sang & De Meulder, 2003). The accuracy of the best-performing method, as measured in leave-one-out validation, is very high: test recall and precision are 94.7% and 95.4% respectively for beginnings of citations, and 98.6% and 97.6% for words inside citations. Subsequently, they matched the extracted in-text references (as well as front-page references) to journal articles in Web of Science (WoS) using rule-based reference parsing. The final dataset consists of all the biotech utility patents granted by USPTO from 2006 to 2010 retrieved from Google Patents, and each patent is linked to a set of its referenced WoS journal articles, in text or on front-page. In total, the dataset for our analysis consists of 33,337 patents and their 860,879 in-text and 637,570 front-page references to scientific papers in WoS.

2.2. Measures

Patent citations. For each patent, we count how many times it is cited by future patents, using a five-year citations time window, following the common practice.

Basicness. For each scientific paper, we measure its basicness using the approach propsod by Weber (2013) for biomedical research, which classifies a paper as highly basic if it only has cell/animal-related MeSH terms but no human-related MeSH terms, moderately basic if it has both cell/animal- and human-related MeSH terms, and not basic (i.e., clinical) if it only has human-related but not cell/animal-related MeSH terms. This measure is an ordinal measure, and its attributes 1, 2, and 3 correspond to not basic, moderately basic, and highly basic, respectively.

Interdisciplinarity we adopt the Rao-Stirling measure (Stirling, 2007), which captures all the three diversity dimensions (i.e., variety, balance, and disparity) of the involved disciplines underlying a study.

Novelty. We adopt the measure developed by Wang et al. (2017), which follows the combinatorial novelty perspective and identifies novel paper as the ones that makes unprecedented combinations of pre-existing knowledge components, where knowledge components are proxied by referenced journals. This measure is a binary variable: 1 if novel and 0 if not novel.

Scientific citations. We count the number of forward citations a scientific paper receives from future papers in the Web of Science (WoS) database, using a five-year citation time window.

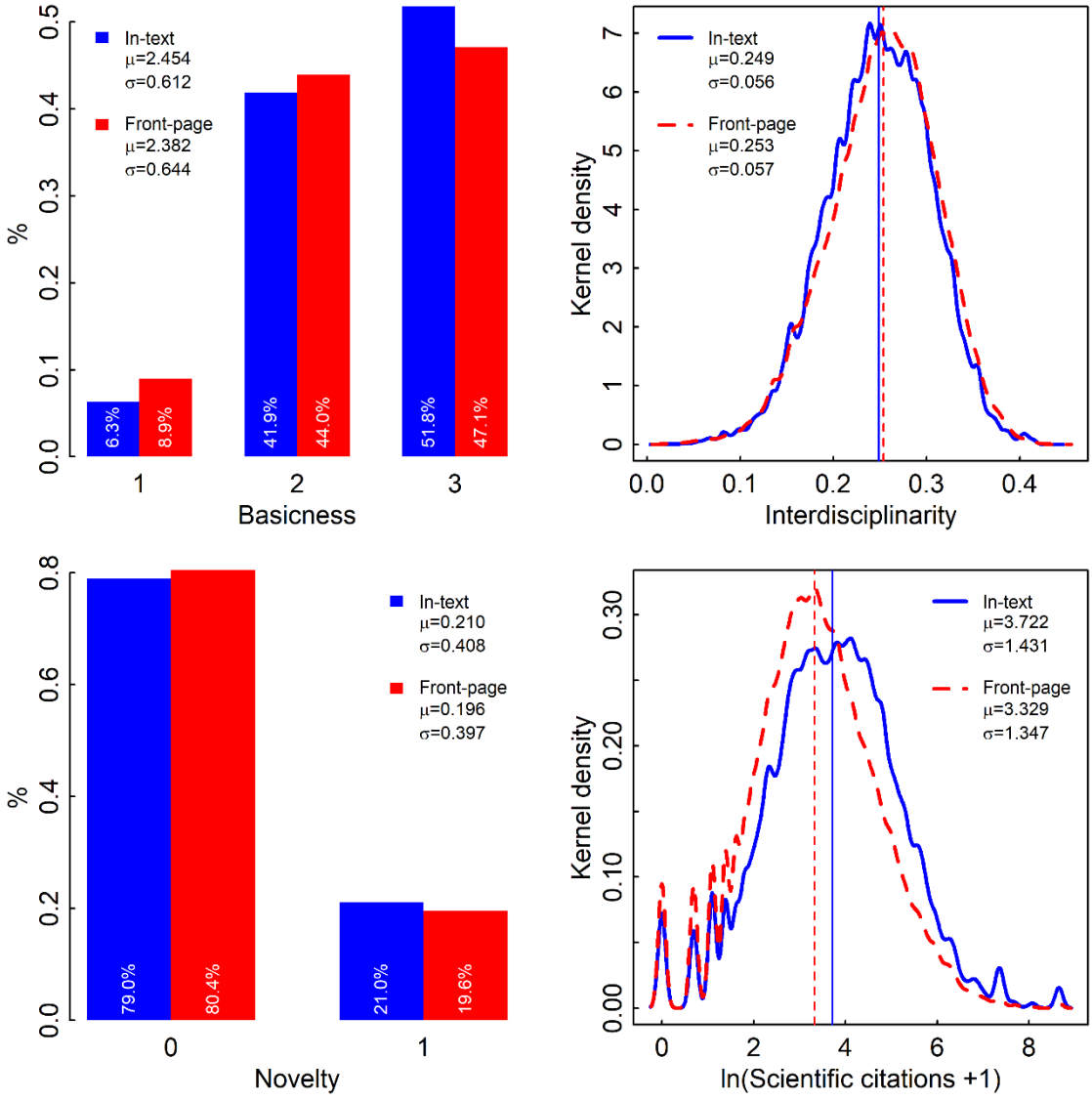
3. Comparing in-text and front-page references

We first examine the overlap between patent front-page and in-text references. The 33,337 USPTO biotech patents in our sample made 1,325,168 references to WoS papers either in text or on front page. In other words, pooling together in-text and front-page reference uncovers 1,325,168 paper-patent-links. Among them 860,879 are in-text references, and 637,570 are front-page references. Figure 1 reports the overlap between in-text and front-page references. In total, 173,281 references appear both in the text and on the front page of the same patent, which accounts for only 20% of all in-text references and 27% of all front-page references.

Figure 1: Overlap between in-text and front-page references.

We further assess the difference between in-text and front-page references in terms of their basicness, interdisciplinarity, novelty, and scientific citations. Figure 2 plots the distributions of these four measures for in-text and front-page references separately. Because the sample size is large, all the mean differences are highly significant (i.e., $p<0.001$) according to Welch two sample t-tests and Wilcoxon rank sum tests, although the difference in interdisciplinarity and novelty seem very small in size. Taken together, results show that in-text references are more basic and have more scientific citations than front-page references. In-text references are less interdisciplinary but more novel than front-page references, but the differences are small. This finding suggests that studies of which kinds of science is more cited by patents might be sensitive to whether the data come from patent in-text or front-page references.

Figure 2. Distribution of basicness, interdisciplinarity, novelty, and scientific citations, by in-text and front-page references.



5. Relationship between science character and patent citations

In the next step we estimate how the characteristics of referenced science affect patent value, as measured by patent forward citations, where referenced science is based on in-text references. The dependent variable is an over-dispersed count variable, so we fit a series of Negative Binomial (NB) models. Regression results are reported in Table 1. Column 1 reports the NB model that uses whether having scientific references as the focal independent variable and incorporates the complete set of patent's issuing year and IPC class dummies. The result suggests that patents having in-text scientific references receive 29.1% more patent citations than patents not having in-text scientific references, issued in the same year and IPC class. Within the set of patents that have in-text scientific references, we further examine the intensity of reliance on science, that is, the number of referenced scientific papers. This independent variable is also a count variable and has a skewed distribution, so we take its natural logarithm for regression analysis. Column 2 shows that as a patent's number of referenced papers increases by 1%, its patent citations increase by 0.122%.

Then we move on to explore the characteristics of referenced science. Column 3-6 each uses average basicness, interdisciplinarity, novelty, and scientific citations of referenced papers as the focal independent variable. In all these models, the ln(number) of scientific references is controlled for, in addition to patent issuing year and IPC class. Avg(Scientific citations) is skewed so it takes natural logarithm transformation for regression analysis. Column 3 shows that, as the average basicness of referenced papers increases by 1, patent citations decrease by 7.0%, holding all other variables fixed. Column 4 suggests no significant effects of interdisciplinarity. Column 5 shows that, as the average novelty of referenced papers increases by 1, patent citations increase by 15.6%, holding all other variables fixed. Column 6 suggests no significant effects of scientific citations. Column 7 further fits a model with all these four variables together and yields consistent results as running separate models for each independent variable (i.e., Column 3-6). In summary, patents building on less basic but more novel science are more impactful in the technological domain.

We repeat all the analyses using front-page references instead, to test whether using front-page will lead to the same findings. As shown in regression results in Table 2. Results are very different.

Table 1. In-text references and patent citations

	Patent citations NB						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
I(sNPR)	0.291***						
	(0.029)						
ln(sNPRs)		0.122***	0.133***	0.122***	0.122***	0.124***	0.136***
		(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)
Avg(Basicness)			-0.070*				-0.073*
			(0.034)				(0.035)
Avg(Interdisciplinarity)				0.622			0.219
				(0.446)			(0.506)
Avg(Novelty)					0.156+		0.175*
					(0.082)		(0.086)
ln(Avg(Scientific citations) +1)						-0.009	-0.009
						(0.013)	(0.014)
Issue year	Y	Y	Y	Y	Y	Y	Y
IPC class	Y	Y	Y	Y	Y	Y	Y
N	33337	26872	26012	26709	26872	26872	25930
BIC	152066	123120	118841	122524	123115	123130	118555

Unit of analysis: patent. The dependent variable is the number of patent forward citations. Each column reports one Negative Binomial regression model. Column 1 estimates the effect of having scientific references. Within the set of patents citing scientific papers, Column 2 estimates the effect of the natural log number of referenced papers. Column 3-6 estimate effects of the average basicness, interdisciplinarity, novelty, and scientific citations separately, controlling for the number of referenced papers. Column 7 fits a model with all these five variables together. All models control for the complete set of patent issuing year and IPC class dummies. Robust standard errors in parentheses. *** $p < .001$, **$p < .01$, *$p < .05$, +$p < .10$.

Table 2. Front-page references and patent citations

	Patent citations NB						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
I(sNPR)	0.229***						
	(0.035)						
ln(sNPRs)		0.210***	0.222***	0.213***	0.210***	0.203***	0.214***
		(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
Avg(Basicness)			-0.007				0.029
			(0.031)				(0.030)
Avg(Interdisciplinarity)				1.759***			2.038***
				(0.356)			(0.393)
Avg(Novelty)					0.102		-0.019
					(0.071)		(0.077)
ln(Avg(Scientific citations)+1)						0.025+	0.051**
						(0.014)	(0.015)
Issue year	Y	Y	Y	Y	Y	Y	Y
IPC class	Y	Y	Y	Y	Y	Y	Y
N	33337	29110	27999	28982	29110	29110	27959
BIC	152144	132262	127270	131696	132269	132267	127043

This table repeats Table 1 but uses science measures based on front-page references instead. Robust standard errors in parentheses. *** p < .001, **p < .01, *p < .05, +p < .10.

3. Conclusion

This paper examined differences between patent in-text and front-page references. Using front-page references cannot replicate the results based on in-text references regarding the relation between the characteristics of referenced science and patent value. Our results also contribute to the studies of patent references to the scientific literature. The inconsistencies between the results based on in-text and front-page references indicate that our results might be sensitive to data source and that we need to be more cautious. Different types of reference might be more appropriate for different research questions.

Open science practices

We have made all computer code underlying the dataset public available. The dataset used for this paper is results of a pilot study. We are currently producing a larger dataset covering all USPTO and EPO patents, and we plan to make the final dataset public.

References

Bryan, K. A., Ozcan, Y., & Sampat, B. (2020). In-text patent citations: A user's guide. *Research Policy, 49*(4), 103946. https://doi.org/https://doi.org/10.1016/j.respol.2020.103946

Callaert, J., Pellens, M., & Van Looy, B. (2014). Sources of inspiration? Making sense of scientific references in patents. *Scientometrics, 98*(3), 1617-1629. https://doi.org/10.1007/s11192-013-1073-x

Hicks, D., Breitzman, A., Sr, Hamilton, K., & Narin, F. (2000). Research excellence and patented innovation. *Science and Public Policy, 27*(5), 310-320. https://doi.org/10.3152/147154300781781805

Ke, Q. (2020a). Interdisciplinary research and technological impact. *arXiv preprint arXiv:2006.15383*.

Ke, Q. (2020b). Technological impact of biomedical research: The role of basicness and novelty. *Research Policy*, *49*(7), 104071. https://doi.org/https://doi.org/10.1016/j.respol.2020.104071

Marx, M., & Fuegi, A. (2020). *Reliance on Science by Inventors: Hybrid Extraction of In-text Patent-to-Article Citations*.

Meyer, M. (2000). What is Special about Patent Citations? Differences between Scientific and Patent Citations. *Scientometrics*, *49*(1), 93. https://doi.org/10.1023/a:1005613325648

Nagaoka, S., & Yamauchi, I. (2015). The Use of Science for Inventions and its Identification: Patent level evidence matched with survey. *Research Institute of Economy, Trade and Industry (RIETI)*.

Narin, F., & Noma, E. (1985). Is technology becoming science? *Scientometrics*, *7*(3), 369-381. https://doi.org/10.1007/BF02017155

Nunn, H., & Oppenheim, C. (1980). *A patent journal citation network on prostaglandins*. Elsevier.

Poege, F., Harhoff, D., Gaessler, F., & Baruffaldi, S. (2019). Science quality and the value of inventions. *Science Advances*, *5*(12), eaay7323. https://doi.org/10.1126/sciadv.aay7323

Popp, D. (2017). From science to technology: The value of knowledge from different energy research institutions. *Research Policy*, *46*(9), 1580-1594. https://doi.org/https://doi.org/10.1016/j.respol.2017.07.011

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora* (pp. 157-176). Springer.

Sang, E. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003,

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society [Journal Article]. *Journal of the Royal Society Interface*, *4*(15), 707-719. https://doi.org/10.1098/rsif.2007.0213

Tamada, S., Naito, Y., Kodama, F., Gemba, K., & Suzuki, J. (2006). Significant difference of dependence upon scientific knowledge among different technologies. *Scientometrics*, *68*, 289-302.

Tijssen, R. J. W., Buter, R. K., & van Leeuwen, T. N. (2000). Technological relevance of science: An assessment of citation linkages between patents and research papers. *Scientometrics*, *47*(2), 389-412. <Go to ISI>://WOS:000089449100014

Verberne, S., Chios, I., & Wang, J. (2019). Extracting and Matching Patent In-text References to Scientific Publications. BIRNDL@ SIGIR,

Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. *Research Policy*, *48*(6), 1362-1372. https://doi.org/https://doi.org/10.1016/j.respol.2019.01.019

Voskuil, K., & Verberne, S. (2021). Improving reference mining in patents with BERT. *arXiv preprint arXiv:2101.01039*.

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

Weber, G. M. (2013). Identifying translational science within the triangle of biomedicine. *Journal of Translational Medicine*, *11*(1), 126. https://doi.org/10.1186/1479-5876-11-126