Delineating the scientific footprint in technology: Identifying scientific publications within non-patent references

Julie Callaert · Joris Grouwels · Bart Van Looy

Received: 29 November 2011 / Published online: 21 December 2011
© Akadémiai Kiadó, Budapest, Hungary 2011

Abstract Indicators based on non-patent references (NPRs) are increasingly being used for measuring and assessing science–technology interactions. But NPRs in patent documents contain noise, as not all of them can be considered ‘scientific’. In this article, we introduce the results of a machine-learning algorithm that allows identifying scientific references in an automated manner. Using the obtained results, we analyze indicators based on NPRs, with a focus on the difference between NPR- and scientific non-patent references-based indicators. Differences between both indicators are significant and dependent on the considered patent system, the applicant country and the technological domain. These results signal the relevancy of delineating scientific references when using NPRs to assess the occurrence and impact of science–technology interactions.

Keywords Science–technology interaction · Non-patent references · Indicators · Machine learning

Mathematics Subject Classification (2000) 68U15

JEL Classification O32 · O34

Introduction

In today’s knowledge based systems of innovation, indicators signaling interactions between scientific and technological activities are highly relevant. Indicators derived from non-patent references (NPRs) within patent documents are very popular in this respect (see e.g. Verbeek et al. 2002). In spite of some discussion about their actual meaning (Nelson 2009), scientific references in patents are in any case indicative of relatedness or closeness

Authors appear in alphabetical order.

J. Callaert (✉) · J. Grouwels · B. Van Looy
ECOOM & Research Division INCENTIM, Faculty of Business and Economics,
K.U. Leuven, Waaistraat 6, Bus 3536, 3000 Leuven, Belgium
e-mail: Julie.Callaert@econ.kuleuven.be

between the developed technology and the cited science (Callaert et al. 2006; Meyer 2000a; Tijssen et al. 2000; Van Looy et al. 2007). The presence of scientific references in the front page section of a patented invention indeed signals the relevancy of these references for assessing and qualifying the claims of the invention. As such, indicators based on scientific references in patents provide useful additional information on science–technology relatedness or vicinity, at least if their presence displays sufficient levels of occurrence (Callaert et al. 2006). Given the widespread and consistent availability of reliable and comprehensive patent databases, indicators based on scientific references in patent documents bear the potential to provide a systematic view on science–technology interactions.

At the same time, NPRs in patent documents contain noise. Several efforts have been made in the past to assess different types of NPRs. Narin and Noma (1985) reported an average of 0.3 NPRs per patent of which about 75% are considered scientific: 48% were journals, 15% were books and 11% were abstracts. Van Vianen et al. (1990) observed that 55.7% of NPRs in Dutch patents were journal citations, the others were mostly books and abstract services. Harhoff et al. (2003) reported similar figures: whereas 60% of NPRs were considered ‘scientific’ references, the remaining 40% referred to trade journals, firm publications or standard texts in technical fields. Callaert et al. (2006), when analysing a sample of 10,000 EPO and USPTO NPRs, found that more than half of them were journal articles. The remaining references included conference proceedings, industry-related documents, and reference books or databases. Guan and He (2007) analyzed Chinese US patents for the period 1995–2004. They found that 70% of all NPRs referred to journal articles or conference proceedings. He and Deng (2007) analyzed a smaller subset of 850 New Zealand USPTO patents, granted between 1976 and 2004, and found that 65% of NPRs referred to the (ISI indexed and non-indexed) scientific literature. The other references referred to company catalogs, manuals, newspapers, gene/plant bank records and patent documents. Lo (2010) analyzed NPRs in US genetic engineering patents, granted between 1980 and 2004. Compared to the other studies, a considerably higher portion of NPRs referred to journals (90%), which is likely due to the specific subfield under study. The remaining references referred to Gene Bank, monographs, technical reports, product catalogs, news items, theses, as well as other patents.

Hence, within the growing literature on science–technology relatedness, and patent based indicators for measuring them, several efforts have been made for further identifying and characterizing NPRs in patents. And even though some non-journal reference categories may still be considered *scientific* in a broader sense, it is clear that not all NPRs are stemming from scientific sources. Therefore, if one is interested in identifying those traces of prior art that refer to scientific research per se—i.e. references to serial peer-reviewed scientific literature (journals) and proceedings—the large scale identification of the scientific character of NPRs becomes highly relevant. The identification methods that have been reported on in the literature are at the same time mostly developed ad hoc, and tailored for the specific application and subset of patents under study. In this article, a method for identification is developed and applied to study the occurrence of actual *scientific* references within NPRs on large scale. We then analyze to what extent NPR-based indicators vary, depending on whether only *scientific* versus all NPRs are taken into account. We consider two NPR-based indicators of the science relatedness of patents. First, the proportion of patents with at least one scientific non-patent reference provides an indication of the ‘extent’ to which science is present within technology. Second, the number of cited scientific NPRs per (citing) patent reflects the science ‘intensity’ of technology. Both indicators are not only compared with their counterparts based on all NPRs: the presence and impact of contingencies—in terms of patent system (EPO vs. US-PTO, technology field and country (national innovation system)—are analyzed as well.

The article is structured as follows: first we describe the methodology and algorithms used to identify scientific non-patent references (SNPRs). Next, we analyze and compare the obtained indicators in terms of occurrence and contingencies. Overall, our observations reveal non-trivial differences for both indicators.

Methodology for characterizing NPRs

A supervised machine learning approach was deployed for classifying NPRs. Machine learning algorithms employ datasets in which every instance (or record) is represented using the same set of features. Supervised machine learning methods start from instances with a priori known labels (i.e. with the corresponding correct outputs). This contrasts with unsupervised learning methods (such as clustering methods), where all instances are unlabelled a priori and the researcher hopes to identify unknown but useful classes from the bottom up.

For our application, where it is known in advance which classes are sought for (namely scientific vs. non-scientific references), supervised machine learning methods are most suitable. In supervised machine learning methods, the goal is to predict class labels based on available features. The resulting classifier is in a next step used to assign class labels to the instances where the values of the predictor features are known, but for which the value of the class label is unknown. This process is called inductive machine learning as a set of rules are derived from instances (a training set) and then applied to the broader population; or—more generally—a classifier is created that can be used to generalize to new instances (Bishop 2006; Kotsiantis 2007; Hastie and Friedman 2009).

In order to arrive at an algorithm that allows delineating scientific references, a ‘learning set’ was created, consisting of 25,783 NPRs. These references were selected randomly from 7,582,096 NPRs, pertaining to all EPO, USPTO and PCT patent documents that are included in the EPO Worldwide Patent Statistical Database PATSTAT (April 2008 version). Each reference in the sample was classified by a team of researchers ($n = 5$) as being either ‘Journal’, ‘Proceedings’ or ‘Non-scientific’ (e.g. manuals, patent abstracts,...). A category ‘Maybe’ was included to designate inconclusive cases (e.g. N.N. (year)). In total 12,465 NPRs or almost 50% received the label ‘Journal’; 2,037 or 8% were classified as ‘Proceedings’; 10,411 NPRs or 41% were non-scientific. A minor 1% (360 NPRs) were doubtful hence labelled ‘Maybe’.

In a next step, all categorized references were parsed, indexed and stemmed in order to create a document by term-matrix that consists of 25,783 rows (references) and 74,127 columns (unique stemmed terms). Cell values equal the frequency of occurrence of each document by term combination.¹ As in most text mining settings, this matrix is very sparse. Contrary to most text-mining applications, no weighting was applied. The main purpose of weighting schemes (e.g. TF-IDF), is to diminish the weight of the terms that occur very often in order to increase the discriminatory power of documents of a more idiosyncratic nature. However, for our particular purpose (i.e. classification of a reference to be scientific or not), words that occur often can be very significant. For example, the two most occurring words in the sample index are “et” and “al”. Their presence in a reference might be a strong indication of the reference being scientific and hence their (potential) impact should not be jeopardized by weighting.

¹ See Salton et al. (1975) on vector space models as well as Magerman et al. (2010) for a more elaborated account on vector space models for patent and publication documents.

In order to assess the robustness of the algorithm, a Monte Carlo-process was used to partition the learning set 10 times in a random way. For each of the partitions, a test set was kept separately for verification purposes (consisting of 30% of the references). The classifier was developed on the references in the partitioned learning sets, excluding the respective test sets. This implies that classifiers were developed on smaller Document-by-Term-matrices, even more so as unique terms were eliminated. This results in 10 different—but partially overlapping—training sets (each consisting of about 17,500 documents and $\pm 56,600$ terms) with their respective test sets of about 8,500 documents. Each one of these training sets generates its own classifier, which means that the process yields 10 different classifiers. If the results of these classifiers on their test sets are similar, it reinforces confidence that any classifier that is generated in this way will yield similar results, also on the rest of the NPR-population.

For each of the 10 training sets, the terms (i.e. all terms that occur in the training set) are ordered by their frequency. Only terms that occur in 10 or more documents over the complete sample of 25,000 references are withheld, as a certain term has to occur frequently enough in order to have ‘predictive’ power for the dataset as a whole. This resulted in 4,148 terms being withheld for the development of the qualifier. Next, a principal component analysis was performed on the training sets. The main purpose of this step is not dimensionality reduction, but de-correlating the variables. Highly correlated variables tend to be linearly dependent, which makes the covariance matrix positive semi-definite, while the discriminant analysis algorithms require positive definite covariance matrices. Retaining 99% of the variance removes the dimensions with an eigenvalue numerically equivalent to zero, solving the problem while retaining almost all information in the dataset. By applying a 99% threshold of withheld variance, 3,500 components are obtained. Next, the obtained eigenvalues were multiplied with the absolute value of the correlation of their respective component with the outcome variable. These scores were used to rank the components. Throughout different simulations, the number of dimensions used to arrive at a classifier (based on discriminant analysis) has been set equal to 10, 50, 100, 500, 1,000 and 3,500 components. The obtained findings reveal that the optimal amount of components for classification purposes amounts to 1,000. When all dimensions are kept, over-fitting occurs as performance on the training sets keeps on rising, while the performance on the test sets starts to diminish.

The classifier is obtained by performing a linear discriminant analysis, where the characterization of the NPR is the result of a linear function of all terms present in the reference. An excerpt of the resulting parameters for this function (top 20 terms in absolute coefficient value) is added in Table 7 of Appendix. Negative coefficients mean that the concerned terms point towards non-scientific references (e.g. ‘genbank’, referring to the genetic sequence database of the NIH; and ‘catalog’, referring to sales or product catalogues). A higher positive coefficient means that the concerned term is more likely to be part of scientific reference (e.g. ‘confer’: referring to conference proceedings; and ‘pna’: referring to the often cited PNAS: Proceedings of the National Academy of Sciences).

The accuracy of the resulting classifier is assessed by calculating the correct classification rates, i.e. the amount of correctly classified NPRs (when comparing to the *a priori* classification that was done by the researchers), divided by the total number of NPRs in the respective sets. The accuracy level, obtained for the training sets, equals 94.1%. For the test sets, an accuracy level of 92% is obtained. Moreover, the different simulations generate highly congruent outcomes, signalling the robustness of the overall approach (Table 1 shows the results for the 10 Monte Carlo runs on the training and test sets).

Table 1 Correct classification rates Monte Carlo runs

Correct classification rate on training	Monte Carlo runs										Average	Std. deviation	
	1	2	3	4	5	6	7	8	9	10			
Number of features kept (by combined correlation and variance measure)	10	0.827	0.826	0.837	0.832	0.830	0.831	0.829	0.838	0.829	0.827	0.831	0.004
	50	0.875	0.875	0.872	0.874	0.875	0.870	0.872	0.874	0.874	0.873	0.873	0.002
	100	0.896	0.891	0.895	0.895	0.895	0.889	0.890	0.894	0.891	0.894	0.893	0.002
	500	0.930	0.928	0.927	0.927	0.926	0.927	0.926	0.926	0.927	0.929	0.927	0.001
	1000	0.941	0.943	0.941	0.940	0.939	0.942	0.941	0.941	0.943	0.941	0.941	0.001
Correct classification rate on test	Monte Carlo runs										Average	Std. Deviation	
	1	2	3	4	5	6	7	8	9	10			
Number of features kept (by combined correlation and variance measure)	10	0.820	0.821	0.839	0.824	0.833	0.835	0.835	0.836	0.835	0.836	0.831	0.007
	50	0.867	0.868	0.874	0.872	0.873	0.880	0.876	0.874	0.873	0.874	0.873	0.004
	100	0.888	0.887	0.893	0.891	0.893	0.895	0.892	0.889	0.891	0.888	0.891	0.003
	500	0.916	0.911	0.920	0.916	0.920	0.919	0.916	0.917	0.918	0.916	0.917	0.002
	1000	0.921	0.917	0.924	0.921	0.925	0.923	0.919	0.921	0.920	0.919	0.921	0.003
3483 (Av.)	0.914	0.913	0.919	0.917	0.920	0.921	0.919	0.919	0.918	0.914	0.912	0.917	0.003

Data

The above-described machine learning methodology was used to characterize all NPRs ($N = 11,388,123$) in the EPO Worldwide Patent Statistical Database PATSTAT (version 04/2009). 58% of all NPRs were characterized as ‘scientific’ (referring to the serial journal literature or to proceedings), implying that approximately 42% of all NPRs are not scientific.

In what follows, we analyze indicators based on NPRs, whereby we are primarily interested in the difference between NPR- and SNPR-based indicators. For these analyses, we consider EPO (applications and grants) and USPTO (grants) patent documents with application years between 2000 and 2009, and with applicant countries in EU15, Switzerland, US, Canada, Japan and Korea. Indicators are broken down by patent system (EPO vs. USPTO), by application year, by applicant country and by technology domain—according to the FhG19 classification; which has been developed by the German Fraunhofer Institute of Systems and Innovation Research, the French patent office and the French Observatoire des Sciences et des Techniques, and which is based on the International Patent Classification (Schmoch 2008). Full counting schemes are used for patents that are assigned to different countries and technology fields. Starting from the number of patents, we calculate for each year/patent system/country/field combination: (1) the average number of patents with NPRs, (2) the average number of NPRs, (3) the *extent* to which science is present in patent documents, measured as the proportion of patents with NPRs and finally (4) the science *intensity* measured as the average number of NPRs per (citing) patent. Each indicator is calculated twice: once using all NPRs and once using only *scientific* NPRs.

Analyses and results

Difference between NPR-based and SNPR-based indicators

Scientific NPRs represent a fraction of all non-patent literature. In Table 2, the differences between NPR- and SNPR-based indicators are shown, and their significance is evaluated using a paired-sample *t* test. Results are presented for EPO and USPTO patents separately, as well as for the aggregate set of EPO and USPTO patents.

Table 2 first reveals some differences between reference-based indicators for EPO and USPTO patents. USPTO patents have considerably higher volumes of references and of patents with references. Also in terms of impact, USPTO patents contain approximately 2.5 to 4 times more (S)NPRs per patent than EPO patents. These observations hold for NPR- as well as SNPR-based indicators and are in line with findings from previous studies. They largely result from different citation requirements and philosophies at the USPTO (duty of disclosure; documentary search) and at the EPO system (no duty of disclosure; patentability search) (Michel and Bettels 2001).

As for the difference between NPR- and SNPR-based indicators, it can be noted that the fraction of NPRs referring to the scientific literature (journal articles and proceedings) equals about 60% of all NPR’s. This leads to considerable differences in derived indicators. The difference between NPR- and SNPR-based indicators is most prominent for the science extent indicator (indicator 3): the proportion of patents containing NPRs is 36% when all NPRs are considered, and is reduced to 19% when only scientific NPRs are considered. These observations are in line with what Guan and He (2007) found for their analyzed subset of Chinese USPTO patents: whereas 30% of patents contained at least one

Table 2 Paired-samples *t* test: NPR- versus SNPR-based indicators (*N* = 6540)

	EPO			USPTO			AGGREGATE		
	All NPRs	Subset SNPRs	<i>t</i> -value (sign.)	All NPRs	Subset SNPRs	<i>t</i> -value (sign.)	All NPRs	Subset SNPRs	<i>t</i> -value (sign.)
(1) Average # patents with references	124.77	67.43	25.39**	201.83	113.67	15.50**	159.33	88.17	25.02**
(2) Average # of references	355.88	219.63	24.26**	1888.80	1110.01	10.77**	1043.35	618.94	12.94**
(3) Average % of patents with references (as portion of total number of patents) → EXTENT	29.78%	15.61%	66.05**	43.80%	22.58%	64.51**	36.07%	18.73%	89.35**
(4) # references/# patents with references → INTENSITY	2.65	2.75	-4.44**	6.84	6.72	1.55 (n.s.)	4.52	4.52	.03 (n.s.)

NPR, only 19% contained at least one scientific NPR. The intensity-indicator measuring the number of references per citing patent (indicator 4) is 4.5 and appears to be independent of whether all NPRs versus only scientific NPRs are considered (except within the EPO subset).²

The observed differences between NPR- and SNPR-based indicators signal the relevancy of singling out scientific NPRs when developing indicators of science–technology relatedness, based on references from patents to the non-patent literature. The difference is especially pronounced for indicators that reflect the extent of science in technology, i.e. the share of patents that contain references to the scientific literature. In the next section, we examine whether these observed differences display distinctive patterns across patent systems, technology fields and (national) innovation systems.

Scientific extent and intensity: patterns of occurrence

Many science–technology studies point out the importance of influencing factors when studying science–technology relations and related indicators (for an overview: see Van Looy et al. 2002). Patent system characteristics, national specificities and technological fields are among the most important factors.

The USPTO and EPO systems differ considerably in terms of search and examination procedures. It has been argued that the comprehensiveness and the quality of citation lists appearing in patent documents vary significantly as a function of the patent office (Meyer 2000b; Michel and Bettels 2001). At USPTO, patent applicants have a duty of disclosure, meaning that they must provide all information that is reasonably deemed necessary to properly examine the patent application and that they were aware of prior to the filing date of the application. When filing for a patent at EPO, applicants are under no such duty of disclosure. As Michel and Bettels (2001) point out, this leads to a situation where the average USPTO search report has the characteristics of a documentary search, whereas the EPO search reports reflect patentability searches. The patentability search is not exhaustive in the same sense as the documentary search in that it should be limited to what is directly relevant to patentability. As such, the volume of references in USPTO patents might be higher than the volume of references in EPO patents, as already confirmed in our dataset of SNPR-based indicators (cf. *supra* Table 2).

Moreover, the relations between science and technology and the indicators for measuring these might be influenced by differences in national innovation systems (Harhoff et al. 2003; He and Deng 2007; Van Looy et al. 2002). The scientific texture that characterizes a country might influence the occurrence of science interactions and hence translate into indicators based on the occurrence of references to the scientific literature.

Finally, technology fields seem important to take into account. As the propensity to publish and the propensity to patent can vary considerably between disciplines and fields, citation patterns might differ within and between scientific and technology domains. Several studies have already pointed at field specific effects regarding indicators that signal science–technology relatedness (Callaert et al. 2006; Guan and He 2007; Harhoff et al. 2003; He and Deng 2007; Van Looy et al. 2002, 2003; Verspagen 2008). Patents from

² This is because the denominators are adapted to the considered subset. Whereas the volume of patents with NPRs is used as the denominator for the NPR intensity, the volume of patents with scientific NPRs is used as a denominator for the SNPR intensity indicator. If, alternatively, the same denominator is used for both (NPR and SNPR) intensities—namely the total number of patents—then both intensities do differ significantly.

technological fields that depend strongly on scientific progress (e.g. pharmaceuticals, biotechnology,...) are expected to display higher proportions and levels of scientific references.

In order to examine to what extent these factors influence SNPR-based indicators, ANCOVA analyses (analysis of co-variance) were performed with scientific extent and intensity (measured by considering only *scientific* NPRs) acting as dependent variables. These variables have been logarithmically transformed to comply with normality assumptions. Independent variables included in the model are the applicant country, the technological domain and the patent system. To account for potential evolutions over time, a year covariate is included in the model. Interaction effects are included between applicant country and technological field. The results are shown in Table 3.

Patent system, technology domain, and applicant country significantly influence both the extent and the intensity of science within technology; as measured by only using scientific NPRs. In addition, one observes a significant interaction effect between technology domain and country.

Figure 1a, b clarify the influence of technological domains on scientific extent and intensity respectively. The figures distinguish between EPO and USPTO patents.

Overall, the observations confirm that the scientific footprint in technology is most outspoken for fields like Pharmaceuticals and Chemicals. At the lower end, one finds domains such as General machinery, Transport, Machine-tools, Energy machinery and Metal products. The technological specificity pattern of the scientific extent in technologies (Fig. 1a) is quite similar for EPO and USPTO patents, although EPO patents show less variety. Also in terms of scientific intensity of technologies (Fig. 1b), the technological specificities are reflected more strongly in USPTO patents, with the science intensity of technology being most notable for pharmaceuticals.

Figure 2a, b show the influence of applicant countries on scientific extent and intensity respectively. Only those countries are included with an annual average of at least 20 patents per technology domain. Again, the figures distinguish between USPTO and EPO patents.

Table 3 ANCOVA analyses: influences on the extent and intensity of science within technology

	DEP VAR = SNPR extent (ln)				DEP VAR = SNPR intensity (ln)			
	Type III sum of squares	df	F	Sig.	Type III sum of squares	Df	F	Sig.
Corrected model	114.648	381	32.028	.000	1037.859	375	16.280	.000
Intercept	.366	1	38.935	.000	.858	1	5.045	.025
Application year	.346	1	36.775	.000	1.120	1	6.588	.010
Patent system	3.544	1	377.173	.000	394.256	1	2319.166	.000
Technology domain	95.464	18	564.497	.000	315.485	18	103.100	.000
Country of applicant	3.149	19	17.640	.000	88.915	19	27.528	.000
Technology domain * country of applicant	9.634	342	2.998	.000	112.069	336	1.962	.000
Error	57.856	6158			835.374	4914		
Total	334.867	6540			13561.553	5290		
Corrected total	172.504	6539			1873.234	5289		
R-square	.665				.554			
Adj R-square	.644				.520

Fig. 1 **a** Influence of technological domains on the extent of science in technology. **b** Influence of technological domains on the intensity of science in technology

Overall, these figures show that the scientific footprint is most prominent in patents from Canada, Belgium, the United States, the United Kingdom and Denmark. For Japan, Austria and Korea, the scientific footprint is least apparent. Country-effects are at the same time shown to depend on which indicator is used. For instance, patents from Belgian applicants appear at the top in terms of extent of science in technology (Fig. 2a), whereas Belgian patents display average levels in terms of scientific intensity (Fig. 2b). The opposite holds for patents from Spanish applicants: although their scientific extent is average, it can be seen that for those Spanish patents that contain scientific references, the

Fig. 2 **a** Influence of applicant countries on the extent of science in technology. **b** Influence of applicant countries on the intensity of science in technology

average number of references is high in comparison to other countries. Finally, the figures reveal that the influence of applicant countries on the scientific footprint in patents is much more outspoken for USPTO patents than for EPO patents.

Influences on differences between NPR- and SNPR-based indicators: altered ranks

We showed earlier that science-technology indicators differ, depending on whether only scientific NPRs or all NPRs are considered (see Table 2). Hence, reference-based indicators are sensitive to the specific subset of NPRs considered when calculating the indicators. The higher this sensitivity, the more appropriate it becomes to distinguish scientific from non-scientific NPRs.

In this section, we investigate to what extent the difference between both types of indicators is influenced by the factors identified in the previous section: patent system, technology domain, and applicant country. ANCOVA analyses are again performed, with the ratios of SNPR- to NPR-based measures as dependent variables. Also these ratios were logarithmically transformed to comply with normality assumptions. The obtained findings are reported in Table 4.

The sensitivity of indicators to the considered subset of NPRs is shown to vary across patent systems, technological domains and applicant countries.

Table 2 (cf. supra) already signaled that the proportion of patents containing SNPRs (extent) is about half the proportion of patents containing (all types of) NPRs. The difference between USPTO and EPO in terms of extent is small in this respect albeit

Table 4 ANCOVA analyses: influences on the ratio between SNPR- and NPR-based indicators

	DEP VAR = Ratio: extent SNPR/ extent NPR (ln)				DEP VAR = Ratio: intensity SNPR/ intensity NPR (ln)			
	Type III sum of squares	Df	F	Sig.	Type III sum of squares	df	F	Sig.
Corrected model	153.157	381	20.322	.000	36.609	375	4.083	.000
Intercept	2.684	1	135.665	.000	.111	1	4.630	.031
Country of applicant	6.337	19	16.860	.000	3.464	19	7.627	.000
Technology domain	111.763	18	313.888	.000	7.674	18	17.833	.000
Patent system	.159	1	8.055	.005	.835	1	34.923	.000
Application year	2.794	1	141.238	.000	.074	1	3.097	.078
Technology domain * Country of applicant	17.378	342	2.569	.000	16.989	336	2.115	.000
Error	110.181	5570			117.480	4914		
Total	1012.880	5952			2687.019	5290		
Corrected total	263.338	5951			154.089	5289		
R ²	.583				.238			
Adj R ²	.553				.179			

significant (ratios of 51.5% vs. 52.4%, overall figures). Similar observations apply for science intensity: while the ratio of both indicators is close to one for both EPO and USPTO, the difference turns out to be significant for the countries and fields under study (for USPTO patents, the intensity indicator shifts from 6.8 to 6.7 while for EPO, the intensity indicator shifts from 2.7 to 2.8³).

As Table 4 reveals, more outspoken and significant differences are observed with respect to countries and especially technology domains. The occurrence of these differences has considerable implications in terms of use and accuracy of indicators. Indeed, the ranking of technology fields—both for scientific extent and intensity—is different, depending on which subset of NPRs is used for developing the indicator. This becomes clear when inspecting Table 5.

While the top 3 technology domains (pharmaceuticals, chemicals and measurement) are ranked consistently, one observes non-trivial shifts in ranks, depending on whether NPRs or SNPRs are considered. This is especially pronounced for the intensity indicator (as witnessed also by the lower rank order correlation): the focus on only scientific references causes the domains of ‘telecommunications’ and ‘computers, office machinery’ to drop 4 places in the ranking, whereas domains of ‘metal products’ and ‘special machinery’ rise 6 and 7 places respectively in terms of scientific intensity.

Similar observations pertain to countries (national innovation systems). Table 6 demonstrates the impact in terms of rankings. Only countries with an annual average of at least 20 patents per technology domain are included.

Table 6 reveals considerable shifts, depending on whether NPRs or SNPRs are used to calculate indicators of science–technology relatedness. While these shifts are most outspoken for the scientific extent indicator (as reflected also in the lower rank order correlation), the resulting changes for the intensity indicator are in any case non-trivial.

³ The intensity indicator for EPO becomes higher because of differential drops in the denominator (number of patents containing SNPRs) and the nominator (number of SNPRs).

Table 5 NPR- versus SNPR-based indicators and ranks across technology domains

Technology domain	Extent NPR (%)	Rank extent NPR	Extent SNPR (%)	Rank extent SNPR	Intensity NPR	Rank intensity NPR	Intensity SNPR	Rank intensity SNPR
Pharmaceuticals	76.6	1	70.0	1	11.64	1	10.30	1
Basic chemicals, paints, soaps, petroleum products	53.1	2	37.8	2	6.57	2	6.47	2
Measurement, control	47.4	3	33.5	3	5.54	3	5.79	3
Electronic components	45.6	4	28.4	4	4.60	5	4.94	7
Telecommunications	42.6	6	26.0	5	4.02	8	3.45	12
Computers, office machinery	44.8	5	25.6	6	4.15	6	3.65	10
Audio-visual electronics	41.8	7	21.8	7	3.43	13	3.02	16
Optics	40.0	8	20.7	8	3.63	10	3.81	9
Medical equipment	30.7	11	18.5	9	4.88	4	5.15	6
Polymers, rubber, man-made fibres	32.6	10	13.5	10	3.48	12	3.84	8
Non-polymer materials	35.4	9	12.8	11	3.76	9	3.64	11
Electrical machinery, apparatus, energy	29.6	12	10.0	12	3.20	14	3.26	14
Textiles, wearing, leather, wood, paper, domestic appliances, furniture, food	25.4	16	9.6	13	4.09	7	5.28	5
General machinery	26.3	13	6.4	14	2.79	15	3.17	15
Special machinery	23.4	18	5.9	15	3.51	11	5.64	4
Machine-tools	25.7	14	4.6	16	2.37	17	2.42	18
Transport	25.7	15	4.1	17	2.46	16	2.35	19
Energy machinery	24.2	17	3.8	18	2.36	18	2.85	17
Metal products	14.9	19	2.7	19	2.29	19	3.39	13
Kendall's tau rank order correlation: 0.833**					Kendall's tau rank order correlation: 0.661**			

Table 6 NPR- versus SNPR-based indicators and ranks across applicant countries

Applicant country	Extent NPR (%)	Rank extent NPR	Extent SNPR (%)	Rank extent SNPR	Intensity NPR	Rank intensity NPR	Intensity SNPR	Rank intensity SNPR
BE	41.1	2	23.3	1	4.37	8	4.85	7
CA	38.5	4	23.0	2	5.92	2	5.87	3
DK	35.4	10	22.5	3	4.85	4	5.37	4
GB	41.1	1	22.3	4	4.70	5	5.32	5
US	36.5	8	20.6	5	6.29	1	6.01	2
NL	38.8	3	20.0	6	3.75	10	3.84	10
FI	36.7	7	18.6	7	3.92	9	3.97	9
CH	37.2	6	18.4	8	4.48	6	4.65	8
FR	37.5	5	18.3	9	3.29	13	3.44	12
SE	35.8	9	17.1	10	4.44	7	5.06	6
ES	29.0	16	16.6	11	5.24	3	7.04	1
DE	35.0	11	16.2	12	3.38	11	3.48	11
IT	32.3	14	15.8	13	2.96	15	3.39	13
KR	31.5	15	15.8	14	2.95	16	3.06	16
AT	32.9	13	15.0	15	3.03	14	3.19	15
JP	34.0	12	14.4	16	3.33	12	3.25	14
Kendall's tau rank order correlation: 0.550**					Kendall's tau rank order correlation: 0.883**			

Conclusion

The analysis presented in this article reveals significant differences between indicators based on all NPRs vis-à-vis indicators relying solely on the subset of SNPRs. This indicates the relevancy of distinguishing scientific references among NPRs if one is interested in assessing the presence and impact of scientific references in relation to technology development. While several scholars already signaled the ‘noisy’ character of NPR-based indicators, this paper is the first to identify—in an exhaustive and accurate manner—scientific references among all NPRs by deploying text mining techniques and machine learning algorithms.

Examining the occurrence of two reference-based indicators of science–technology relatedness (the proportion of patents containing at least one (scientific) NPR: extent of scientific presence; and the average number of (scientific) NPRs per (cited) patent: science intensity of the patent) reveals contingencies in terms of patent system, technology domain and country (national innovation system). In addition, these contingencies influence the actual ratio of SNPRs. These observations further corroborate the relevancy of delineating scientific references when constructing science–technology indicators based on NPRs.

Acknowledgments This article is an extended version of a paper presented at the 13th International Conference on Scientometrics and Informetrics, Durban (South Africa), 4–7 July 2011 Callaert et al. 2011). The authors want to acknowledge conference participants who contributed with comments and remarks.

Appendix

See Table 7.

Table 7 Parameters linear discriminant function (excerpt: top 20 terms in absolute coefficient value)

Constant Term	1.3014 Coefficient
Workshop	5.9842
Symposium	5.1527
Meet	4.4552
Abst	−4.4254
Disclosur	−4.2663
Confer	4.1981
Pna	4.1401
Genbank	−3.9761
Chapter	−3.7306
Handbook	−3.6599
Natur	3.5998
Transact	3.5579
Manual	−3.4234
Catalog	−3.4091
Acta	3.4032
Journal	3.3897
Proc	3.2414
Lett	3.1426
Annal	3.0593
Tran	3.0394

References

- Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
- Callaert, J., Grouwels, J., & Van Looy, B. (2011). Delineating the scientific footprint in technology: Identifying scientific publications within non-patent references. In: E. Noyons, P. Ngulube & J. Leta (Eds.), *Proceedings of ISSI 2011—The 13th International Conference on Scientometrics and Informetrics, Durban* (pp. 13–18). San Juan: ISSI.
- Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics*, 69(1), 3–20.
- Guan, J. C., & He, Y. (2007). Patent-bibliometric analysis on the Chinese science—Technology linkages. *Scientometrics*, 72(3), 403–425.
- Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8), 1343–1363.
- Hastie, T., & Friedman, J. (2009). *The elements of statistical learning—Data mining, inference, and prediction* (2nd ed.). New York: Springer-Verlag.
- He, Z. L., & Deng, M. (2007). The evidence of systematic noise in non-patent references: A study of New Zealand companies' patents. *Scientometrics*, 72(1), 149–166.
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In I. E. Magogiannis, et al. (Eds.), *Emerging Artificial Intelligence Applications in Computer Engineering*. Amsterdam: IOS Press.
- Lo, S. C. S. (2010). Scientific linkage of science research and technology development: A case of genetic engineering research. *Scientometrics*, 82(1), 109–120.
- Magerman, T., Van Looy, B., & Song, X. (2010). Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2), 289–306.
- Meyer, M. (2000a). Does science push technology? Patents citing scientific literature. *Research Policy*, 29, 409–434.

- Meyer, M. (2000b). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, 49(1), 93–123.
- Michel, J., & Bettels, B. (2001). Patent citation analysis: A closer look at the basic input data from patent search reports. *Scientometrics*, 51(1), 185–201.
- Narin, F., & Noma, E. (1985). Is technology becoming science? *Scientometrics*, 7, 369–381.
- Nelson, A. J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy*, 38, 994–1005.
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620.
- Schmoch, U. (2008). Concept of a technology classification for country comparisons. Final report to the world intellectual property organization.
- Tijssen, R. J. W., Buter, R. K., & Van Leeuwen, T. N. (2000). Technological relevance of science: Validation and analysis of citation linkages between patents and research papers. *Scientometrics*, 47, 389–412.
- Van Looy, B., Callaert, J., Debackere, K., & Verbeek, A. (2002). Patent-related indicators for assessing knowledge-generating institutions: Towards a contextualised approach. *The Journal of Technology Transfer*, 28(1), 53–61.
- Van Looy, B., Magerman, T., & Debackere, K. (2007). Developing technology in the vicinity of science: An examination of the relationship between science intensity (of patents) and technological productivity within the field of biotechnology. *Scientometrics*, 70(2), 441–458.
- Van Looy, B., Zimmermann, E., Veugelers, R., Verbeek, A., Mello, J., & Debackere, K. (2003). Do science–technology interactions pay off when developing technology? *Scientometrics*, 57(3), 355–367.
- Van Vianen, B., Moed, H., & Van Raan, A. (1990). An exploration of the science base of recent technology. *Research Policy*, 19, 61–81.
- Verbeek, A., Debackere, K., Luwel, M., & Zimmermann, E. (2002). Measuring progress and evolution in science and technology-I: The multiple uses of bibliometric indicators. *International Journal of Management Reviews*, 4(2), 179–211.
- Verspagen, B. (2008). Knowledge flows, patent citations and the impact of science on technology. *Economic Systems Research*, 20(4), 266–339.