Patent Citation Analysis with Google[1]

Kayvan Kousha and Mike Thelwall

Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.

E-mail: {k.kousha, m.thelwall}@wlv.ac.uk

Citations from patents to scientific publications provide useful evidence about the commercial impact of academic research but automatically searchable databases are needed to exploit this connection for large scale patent citation evaluations. Google covers multiple different international patent office databases but does not index patent citations or allow automatic searches. In response, this article introduces a semi-automatic indirect method via Bing to extract and filter patent citations from Google to academic papers with an overall precision of 98%. The method was evaluated with 322,192 science and engineering Scopus articles from every second year during 1996-2012. Although manual Google Patent searches give more results, especially for articles with many patent citations, the difference is not large enough to be a major problem. Within Biomedical Engineering, Biotechnology, and Pharmacology & Pharmaceutics, 7%-10% of Scopus articles had at least one patent citation but other fields had far fewer so patent citation analysis is only relevant for a minority of publications. Low but positive correlations between Google Patent citations and Scopus citations across all fields suggests that traditional citation counts cannot substitute for patent citations when evaluating research, however.

Introduction

Bibliometric methods are commonly used to help assess the research impact of scientific publications based upon citations in conventional citation indexes, such as the Web of Science (WoS) and Scopus. However, some scholarly publications have a commercial utility that does not directly translate into academic citations. Since governments value research that has commercial value, alternative methods are needed to assess it in order to track it or adequately reward its authors. Citations to academic publications from patents, for instance, suggest the commercial value or at least technological innovation of the cited article. In some cases, the invention may even have been triggered by the academic research (Verbeek, et al., 2002). Citations from patents have been used to assess the relationship between science and industry (e.g., Narin & Olivastro, 1992; Schmoch, 1993; Narin, Hamilton, & Olivastro, 1997). Of the many bibliometric studies using patent citations (e.g., Tijssen, Buter, & Van Leeuwen, 2000; Tijssen, 2001; Callaert et al., 2006; Meyer, Debackere, & Glänzel, 2010; Callaert, Grouwels, & van Looy, 2012; Roach & Cohen, 2013), some applications include assessing the technological value of academic journals (Huang, Huang, & Chen, 2014; Liaw, Chan, Fan, & Chiang, 2014), the research performance of firms (Nagaoka, 2007; Subramanian & Soh, 2010; Hung, 2012), university-industry knowledge relationships (Leydesdorff, 2004) and the research performance of countries (Van Looy et al., 2003). Hence patent citations are an important data source for bibliometrics but, as discussed below, they are difficult to use for large scale studies.

[1] This is a preprint of an article to be published in the Journal of the Association for Information Science and Technology © copyright 2015 John Wiley & Sons, Inc.

Bibliometric studies of patent citations can use time-consuming manual searches of patent databases, such as Google Patents or the Derwent World Patents Index, none of which currently include an academic citation index to aid the process. This is difficult for researchers or evaluators who want to estimate the overall citation impact of large numbers of articles from a reasonably comprehensive collection of patents (Verbeek et al., 2002; Shirabe, 2014) and so many studies are restricted to patents from a narrow range of years and disciplines (e.g., Tijssen, Buter & van Leeuwen, 2000; Meyer, 2003; Callaert et al., 2006; Meyer, Debackere, & Glänzel, 2010; Callaert, Grouwels, & van Looy, 2012). Although attempts have been made to partially automate this process (e.g., Lawson, Kemp, Lynch, & Chowdhury, 1996; Nanba, Anzen, & Okumura, 2008; Lopez, 2010; Ma, Sun, Wang, & Yang, 2010; Callaert, Grouwels, & van Looy, 2012), these methods require all the patents to be downloaded or collected as a first step, which is inefficient and requires repeated downloading to keep the results up-to-date.

In response to the above problems, this article introduces a practical method to help evaluators and funders to extract patent citation counts from Google Patents for large collections of academic articles. Google Patents contains a large collection of fully searchable patents from the United States Patent and Trademark Office (USPTO) since 1790 and the European Patent Office (EPO) since 1978 (https://support.google.com/faqs/answer/2539193?). It also indexes patents from the World Intellectual Property Organization (WIPO), Canada, China and Germany. Google Patents is not a citation index but its full-text search capability can be used to locate citations to other scientific publications in the patent references. It does not support automatic API searches (https://developers.google.com/patent-search/) and its main search interface is impractical for large-scale research evaluations. To solve these problems, the new method introduced here uses automatic Bing searches (exploiting Bing's crawl of the Google Patents website) in combination with automatic duplicate results filtering of the data returned by Bing. The method was evaluated with a study of citations to 322,192 Scopus articles across sixteen science and engineering fields and manual checks of a sample of the search results.

Patent Citations

Three decades ago, Narin and Noma (1985) first convincingly argued that patent citations to scientific papers could be used to investigate the relationship between science and technology. They found a significant number of citations from US biotechnology patents 1978-1980 to scientific publications. About half (48%) of the non-patent references (i.e., patent citations to documents other than other patents) were to journal articles. A slightly larger proportion (56%) was found later for Netherlands-invented U.S. patents between 1982 and 1985 in all technological fields (van Vianen, Moed, & van Raan, 1990) and so it is likely that academic research is the most important non-patent source of evidence for U.S. patents. A large-scale study analysed 430,226 non-patent references in about 397,600 U.S. patents issued 1987-1988 and 1993-1994 (Narin, Hamilton, & Olivastro, 1997). For 1993-1994 patents there were 1.5 non-patent references per patent. About 73% of the papers cited by industry patents were public science, authored at academic, governmental, and other public institutions, with the remaining 27% authored by industrial scientists. There was a strong tendency for inventors to cite articles authored in their own country, at prestigious universities and laboratories, and supported by well-known founding bodies such as the National Institute of Health (NIH) and The National Science Foundation (NSF). This confirms that public science plays an important role in supporting U.S. industry and technology. Instructions in the US for applicants to include a complete list of references to the state of the art have led to 3.5 times more non-patent references than in European patents, however (Michel & Bettels, 2001), and so other patent databases may contain substantially fewer academic citations.

Out of 33,127 pharmaceutical EPO patents 1990-1997, 46.5% of the non-patent references were to scientific articles in ISI (Institute for Scientific Information) databases (now WoS) (Brusoni, Criscuolo, & Geuna, 2005). A study of 10,000 patents from 1991-2001 found that 55% (USPTO) and 64% (EPO) of the non-patent references were to journal articles (Callaert et al., 2006). Similarly, in Chinese-authored USPTO patents 1995–2004, 64% of the non-patent references were to journal articles (Guan & He, 2007) and in New Zealand USPTO patents 1976-2004, 65% of the non-patent references were to either WoS-indexed journals (52.6%) or other articles (12.6%) (He & Deng, 2007). In contrast, 90% of the non-patent references were to journal articles in study of 6,274 USPTO genetics patents 1980–2004 (Lo, 2010), revealing the importance of disciplinary differences. Overall, however, the majority of non-patent references in patents are probably to scientific articles.

A study of patent citations to Dutch research papers from U.S. patents 1987-1996 found a low but statistically significant correlation (Pearson r=0.16, n=2,241) between academic citations and USPTO citations to Dutch papers from 1993-1996, suggesting that academic and commercial impact have something in common (Tijssen, Buter, & van Leeuwen, 2000). Supporting this, although nearly all respondents (94%) to a survey of 35 Dutch inventors believed that their internal research was important or critical for their patents, non-patent citations were important sources of information in 70% of cases. The majority of these non-patent citations were created by the applicants themselves rather than by examiners (Tijssen, Buter, & van Leeuwen, 2000). Applicants were also found to create more references than did examiners for 502,687 patents issued by USPTO during 2001-2003, with the difference being 26% (Sampat, 2004).

Citations in patents can reflect the differing citing motivations of both patent authors and examiners (Schmoch, 1993; Meyer, 2000a; Oppenheim, 2000) and some do not reflect technological innovation (Jaffe, Trajtenberg, & Fogarty, 2000), although this problem is less frequent in self-citations (Li et al., 2014). Citations from patent examiners probably do not reflect knowledge flows from public research to industry (Alcacer & Gittelman 2006; Alcácer, Gittelman, & Sampat, 2009; Roach & Cohen, 2013), but this is not a substantial problem because they are in a minority (Tijssen, Buter, & Van Leeuwen, 2000; Sampat, 2004; Lemley & Sampat 2010). The share of scientific references in patents seems to differ across technological domains (Callaert et al. 2006), patent offices (Michel & Bettels, 2001) and between domestic and international patents (Tijssen, 2001) and so results from one investigation may not be valid in other contexts. Overall, however, scientific references in patents can be used cautiously as indicators of the impact of science on technology (e.g., Meyer, 2000b; Callaert, Pellens, & Van Looy, 2014).

A study of the relationship between patent citations and citation impact for articles from the Science Citation Index (SCI) matched nanoscience and nanotechnology SCI articles 1991–2004 with patent citations in the Derwent Innovations Index (DII) up to 2004. Few (4.6% or 7,000) of the SCI papers had received at least one patent citation and less than 1.5% (2,000) had been cited at least twice. However, about 14% of the most cited papers had also been cited in patents, indicating that highly cited articles are more likely to receive patent citations in this field (Meyer, Debackere, & Glänzel, 2010).

Non-patent citations have been used to indicate the technological value of academic journals in a similar way to Journal Impact Factors (JIFs) for scientific impact (Huang, Huang, & Chen, 2014; Liaw, Chan, Fan, & Chiang, 2014). One study matched the references in 2011 US patents with journal articles published in journals in the 2011 Journal Citation Reports (JCR) to generate five and ten-year Technological Impact Factors (TIFs). There were low positive correlations between five and ten-year TIFs and JIFs, suggesting that TIFs may reflect a genuinely different type of impact (Huang, Huang, & Chen, 2014).

Text mining techniques to extract information from patents, such as with keywords from titles or abstracts, can be used for patent analyses (e.g., Yoon & Park, 2004; Tseng, Lin, & Lin, 2007; Lee, Yoon, & Park, 2009). Several studies have used keywords in patent titles to analyse technical topics (Han et al., 2014), technological trends (Courtial, Callon, & Sigogneau, 1993) or university–industry relationships (Leydesdorff, 2004).

Patents play a more significant role in some technological fields than in others (see Cohen, 2010). According to World Intellectual Property Organization (WIPO) statistics, the share of US patent applications between 1999 – 2013 is higher in technology fields such as Computer Technology (10.8%), Medical Technology (7.9%) and Pharmaceuticals (6.3%) than in other fields (http://www.wipo.int/ipstats/en/statistics/country_profile/profile.jsp?code=US) Another issue is that there might be some areas that are frequently patented but these patents are rarely informed by scientific research. In contrast, Biotechnology, biomedical science and pharmaceutics have a high linkage with scientific research, as discussed above (see Narin & Olivastro 1992; Narin, Hamilton, & Olivastro, 1997; Verbeek et al. 2002).

Experiments in Web Citation Extraction

Web-based citation analyses of academic publications have the potential to provide data about the wider impact of research beyond that of conventional citation indicators (Cronin, 2001). For example, webometric investigations have attempted to extract impact evidence from scholarly articles, digitised books, clinical guidelines, online presentations, and academic course syllabi on the web.

The Web: Counting citations to academic articles from all web pages provides a free impact indicator that correlates with conventional citations at both the article and journal levels (Vaughan & Shaw, 2003, 2005). Most of these web citations originate from web CVs, journal tables of contents and library web sites, which are mainly created for navigational, self-publicity or current awareness (Kousha & Thelwall, 2007b). In response, later studies extracted web citations from specific parts of the web in order to get more targeted information for the impact assessment.

Google Scholar: Google Scholar indexes citations from documents that are online or provided by publishers. It seems to cover about 88% (100 million) of the English-language scholarly documents accessible on the web (Khabsa & Giles, 2014), which is double the size of WoS (about 53 million: authors' own WoS searches in March 2015). Google Scholar indexes more citations than do WoS and Scopus in many fields, especially in the social sciences, arts and humanities and computing (Meho & Yang 2007; Kousha & Thelwall, 2007a; Bar-Ilan, 2008; Kulkarni, Aziz, Shams, & Busse, 2009). Although the *Publish or Perish* software can be used to automate data gathering from Google Scholar in some extent (Harzing & van der Wal, 2008), it cannot be used to extract citations from Google patents. The substantial coverage of Google Scholar seem to be useful for assessments of recently-published, in press publications and other publications (e.g., non-English) which have been invisible in conventional citation indexes. Hence, the derived citation indicators from Google Scholar such as h-index found to be much higher than either WoS or Scopus (e.g., Amara & Landry, 2012; De Groote & Raszewski, 2012). Moreover, the retroactive growth of Google Scholar citation seems to be considerably higher than WoS, making it a promising tool for citation tracking (de Winter, Zadpoor, & Dodou, 2014). Nonetheless, the lack of quality control and potential for manipulation of citation results makes it problematic to use Google Scholar for impact assessments (e.g., Jacso 2011; Beel & Gipp, 2010; López-Cózar, Robinson-García, & Torres-Salinas, 2014). Google Scholar includes citations from patents but extensive manual searching and filtering are needed to locate them and automatic searching is not allowed for this.

Google Books: Google Books does not index citations but contains digitised versions of millions of books and can be searched for citations (Kousha & Thelwall, 2009). Google Books citations are useful for the impact assessment of research, especially in book-based fields, because existing citation indexes include few citations from books (Kousha, Thelwall, & Rezaie, 2011). Although the Google Books API (Applications Programming Interface) can be used to automate citation counting from digitised books (Kousha & Thelwall, 2014) and it uses similar scanning technology to Google Patents, Google Books searches do not include citations from patents.

Online Presentations: Citations from online presentations can give impact evidence in conference-based subject areas, such as computer science and engineering, where proceedings papers are important. Many conference papers have associated presentation files (e.g., in Microsoft PowerPoint) that are shared online (e.g., in slideshare.net or slideshow.com). Citations from presentations can be collected by automatic Bing searches with web queries that combine bibliographic information with the advanced search operator filetype:ppt to restrict the results to presentation files (Thelwall & Kousha, 2008).

Clinical Guidelines: Citations from clinical guidelines directly reflect the impact of published research on the treatment of patients and are to some extent the health equivalent of patent citations. These citations can sometimes be systematically gathered from websites that publish them, such as the National Institute of Health and Clinical Excellence (NICE) site in the UK. A study of NICE citations using this method found that articles cited in guidelines are more likely to be highly cited in WoS (Thelwall & Maflahi, 2015).

Academic Syllabi: Mentions of academic outputs in course reading lists can be used as an indicator of their value for teaching utility of publications (Kousha & Thelwall, 2008) and this is particularly useful for textbooks and introductory science books that have primarily educational value. It is possible to automatically count mentions of monographs in online academic syllabi through a combination of Bing searches and rules to filter out false matches. Over a third of 14,000 monographs in one study had at least one academic syllabus mention, with more in the arts and humanities (56%) and social sciences (52%), confirming the importance of monographs for teaching in book-based subject areas (Kousha & Thelwall, in press).

In summary, there is empirical evidence that different types of citations can be extracted from the web for impact assessment. Nevertheless, no previous study has used an automatic method to extract large numbers of patent citations from the web.

Research Questions

This study introduces and assesses a new method to semi-automatically extract patent citations on a large scale from the web. The technique exploits both the Google Patents database and the Bing API automatic search interface. The following questions drive the evaluation and investigations into the value of the extracted information.

1. Can citations to academic articles be automatically extracted from the Google Patent database with an acceptable degree of coverage and accuracy?
2. Do Google Patent citations correlate with Scopus citations to academic articles?
3. How do publication date and discipline affect the answers to the above questions?
4. Which patent offices do the Google Patent results mainly originate from?

5. Are the topic areas that tend to attract Google Patent citations the same as the topic areas that tend to attract Scopus citations?

Methods

The research design is to use the new method to extract Google Patent citations to a large number of academic articles from multiple fields, and then to assess a random sample for coverage and accuracy (RQ1), to correlate the results against citation counts separately by field (RQ2), to analyse the main results by year and discipline (RQ3). The patent office origins were checked separately (RQ4) and a method introduced to analyse keywords (RQ5).

Data Sets

Bibliographic information and citation counts for English language articles from every second year 1996-2012 were extracted from Scopus from sixteen science and engineering fields (see Table 3). These years were selected to investigate the impact of time on citations to relatively recent academic articles in patents. Older articles are unlikely to be relevant to research evaluations and newer articles are unlikely to have attracted patent citations, given the time delay between submitting a patent and getting it approved. The sixteen science and engineering fields were selected to represent a range of different subject areas that seem likely to be useful for patents and so represent an estimated best case scenario for patent citation analysis.

The search command *subjmain* was used in the advance Scopus search interface to retrieve publications in a specific field (e.g., subjmain(1303) for Biochemistry). The results were limited to articles (excluding books, editorial materials, letters, notes, short surveys and reviews) within the selected subject areas. English language articles only were selected because Google Patents largely covers US and European patents in English and citations to non-English articles might be underrepresented in the results. The inclusion of other patents might therefore have biased the findings. Elsevier's Scopus was chosen instead of WoS due to its greater coverage of international academic publications (e.g., 20,000 peer-reviewed journals) and its feature to export citation information for large numbers of articles (20,000 records via email), enabling a large scale evaluation.

For each selected field and year, a random sample of 2,250 articles was taken from the Scopus set (e.g., Biotechnology articles published in 2002). Thus, 20,250 Scopus records were initially gathered for each field (2,250 * 9 years = 20,250). Cases where the Scopus bibliographic information was too incomplete to generate effective patent searches were removed from the samples. For instance, some Scopus records lacked author information, giving “Anon” or “Dull” instead. Scopus articles with less than three words in their titles were also excluded. Some of these were non-article documents (e.g., Preface, Clinical quiz, Technology summary) that apparently were incorrectly labelled as articles in Scopus. Moreover, very short titles can generate false search matches (e.g., “Dietary fiber”, “Material developments”, “Trace analysis”) even when they are combined with authors names (see also the limitations). Thus, a few records were excluded from the initial sample of articles, ranging from 55 in Pharmacology & Pharmaceutics to 204 in Control & Systems Engineering (Table 3).

Automatic Google Patent Searches with Bing

Bing was used for automatic searches because it is the only major search engine that supports API searches and initial testing suggested that it indexes much of Google Patents. The free software *Webometric Analyst* (http://lexiurl.wlv.ac.uk) was used to perform automatic searches with the Bing API. The queries were automatically generated from the Scopus bibliographic information using an application exclusively designed in Webometric Analyst for this purpose (see “Make Google Patent

Searches for Scopus/WoS Data" option in "Make Searches" menu): the first author last name, the first (up to) ten terms of the article title as a phrase search, and the publication year. The *site:google.com/patents* command was added to each query to limit the results to the Google Patents website, as the examples below show.

- *Shim "Transdermal delivery of mixnoxidil with block copolymer nanoparticles" 2004 site:google.com/patents*
- *Doldi "Intragastric balloon in obese patients" 2000 site:google.com/patents*
- *Vidic "Uptake of elemental mercury vapors by activated carbons" 1996 site:google.com/patents*

Journal names were not included in the queries because a variety of abbreviations are used in patents (e.g., "Mol. Plant-Microbe Interact." or "Mol Plant Microbe In") and this would reduce the number of correct matches. Microsoft allows 5,000 free queries per month for the Bing API but additional queries can be purchased from Microsoft and used in Webometric Analyst.

Removing Duplicate Matches

In a few cases the Bing search results included multiple versions of the same or overlapping patents, such as the initially submitted version and the finally accepted one, or an original patent and a continuation or continuation-in-part. In order to avoid counting both, we filtered out patents with the same titles and descriptions in the Bing API search results. For instance, we found three citation matches from the same query with the identical condensed titles "Nerve stimulation for treating ..." (first column in Table 1) in the Bing results. However, only two matches were unique and one was a duplicate. Bing shortens patent titles in its results and therefore omitting search matches based on titles alone (after "- "in the title field) may exclude relevant matches (e.g., the third row in Table 1). Hence, as a practical method a filtering capability was added to Webometric Analyst to exclude duplicate matches based upon both titles and descriptions in the Bing results (see the Utility option in Webometric Analyst "Filter out duplicate title and description for the same query"). This removed about 3% (1,564 out of 52,453) of the initial Bing results searches, ranging from 1.4% Biochemistry & Molecular Biology to 4.6% Electrical & Electronic Engineering. This shows that the raw results from Bing API search to locate citations in Google Patents include duplicates (see Table 2) but suggests that they a low proportion of the results.

Table 1. An example of non-duplicate and duplicate Bing API results.

Title from the Bing API search*	**Text description from the Bing API search**	**Patent URL**	**Complete patent title (first inventor's name)**
Patent US6892098 - Nerve stimulation for treating ...	Deurloo K E et al., "Transverse tripolar stimulation of peripheral nerve: a modelling study of spatial selectivity," Med Biol Eng Comput, 36(1) ... 1998, which is ...	google.com/patents/US6892098	Nerve stimulation for treating spasticity, tremor, muscle weakness, and other motor disorders (Shai Ayal)
Patent US7324853 - Nerve stimulation for treating ...	Deurloo K E et al., "Transverse tripolar stimulation of peripheral nerve: a modelling study of spatial selectivity," Med Biol Eng Comput, 36(1) ... 1998, which is ...	google.com/patents/US7324853	Nerve stimulation for treating spasticity, tremor, muscle weakness, and other motor disorders (Shai Ayal)
Patent US7885709 - Nerve stimulation for treating ...	Nerve stimulation for treating disorders ... A number of patents describe ... "Transvers e tripolar stimulation of peripheral nerve: a modelling study of …	google.com/patents/US7885709	Nerve stimulation for treating disorders (Tamir Ben-David)

**Deurloo 1998 "Transverse tripolar stimulation of peripheral nerve: a modelling study" site:google.com/patentsbelow patents*

Manual Checks of the Bing API Searches

Manually checking was used to investigate the coverage of the filtered automatic Bing searches. A stratified sample of 320 results was selected for this with five high, medium, low and uncited articles in the online patents from the original Bing API searches for each field (20 results * 16 fields = 320). Google Patent searches sometimes retrieve many results with similar titles, abstracts and inventors from different patent applications and different international patent offices (e.g., Google Patents: US20100298403, EP1904436, WO2007006700, CA2612211, CN101218204). To check for this, each article was searched for in the Google Patent main search interface and duplicate patents were identified by comparing the patent titles, authors and initial descriptions, but ignoring the patent number. Patents were considered to be duplicates if they described the same invention, even if the wording was slightly different because these changes were presumably due to revisions of the text rather than major changes to the product.

Results and Discussion

Estimating the Accuracy and Coverage of Bing API Searches

The manual checking of the stratified sample of 320 articles (Uncited, low-cited, medium-cited and high-cited) found that the Bing API results tended to be less comprehensive than the direct manual Google Patent searches. For uncited articles, we randomly selected five articles without citation in the Bing API search results across sixteen fields. For low-cited articles, we randomly chose five articles with one or two citations in the Bing API searches. For medium-cited articles, we calculated the median for each field after excluding articles with zero, one, two and five highly cited articles (see below) and used it as an indicator to select medium-cited articles. For instance, in Food Science, Environmental Science, Industrial & Manufacturing Engineering and Pharmacology & Pharmaceutics the medians were 4, but in Biotechnology and Biomedical Engineering the medians were 5 and 6 respectively. For highly-cited articles, we selected top five highly cited articles in each field in order to check the extreme citation results from the Bing against Google manual patent searches. For instance, in Biomedical Engineering we selected articles with 139, 71, 69, 62 and 61 citations from the Bing API search results. Out of 80 uncited articles from the Bing API searches, 8(10%) had (1 or 2) citations in the manual Google Patent searches. For instance, no results were found for the Bing API query *Hsieh "Isolation and characterization of a functional A-type cyclin from maize" 1998 site:google.com/patents*, whereas Google Patents retrieved two correct citations. Since 90% or more of the articles in each of the sixteen fields had no patent citations as returned by the Bing API searches, the uncited articles form the vast majority of the results and so the automatic searches give the same results as manual searches nearly all of the time. The situation is not the same for the more important cases where there are some patent matches, however.

Google Patents also returned more correct matches than did the Bing API searches for the other three samples (Table 2). For example, the query *Barga "Recovery guarantees for internet applications" 2004* in the Google Patents main search interface retrieved eight different results, only one of which was also present in the Bing searches. However, as shown in Table 2, about a third (31.2%) of the Google Patent manual search results were duplicates (at least from a citation counting perspective). In a few cases, Bing retrieved patent citations not returned by manual Google Patent searches, perhaps because Google Patents does not display all relevant results when there are many matches. As an example, the query *Peng "Role of polymers in improving the results of stenting in" 1996 site:google.com/patents* correctly matched 60 non-duplicate patents in Bing that were not in the Google Patents manual search results (e.g., US7794743, US8828418, US7285304, US7622070, US6986899).

For each sample, the recall statistic from information retrieval can be estimated. This measures the percentage of the full set of matches found by any given search. It can be estimated by dividing the total number of Bing citations by the total of the combined Bing and manual unique results, as follows.

- Uncited in Bing: Recall = 0/(0+10) or 0%.
- Low cited in Bing: Recall = 126/(0+191) or 66%.
- Medium cited in Bing: Recall = 456/(0+677) or 67%.
- High cited in Bing: Recall = 3208/(273+3510) or 85%.

Although based upon low numbers of articles and hence unreliable, these figures suggest that the semi-automatic method tends to get at least two thirds of the citations to articles that it finds at least one citation for.

Table 2. A comparison of the automatically filtered Bing API results and the Google Patent (GP) manual searches.

Sample	Articles checked	Bing citations* (Median)	All GP citations + (Median)	GP non-duplicate citations++ (Median)	GP results not found by Bing	Bing results not found by GP
Uncited in Bing	80	0 (0)	11 (0)	10 (0)	10	0
Low cited in Bing	80	126 (2)	279 (3)	191 (2)	65	0
Medium cited in Bing	80	456 (5)	1,034 (11)	677 (8)	221	0
High cited in Bing	80	3,208 (33.5)	5,054 (66.5)	3,510 (40.5)	302	273
Total	320	3,790 (2.5)	6,378 (6)	4,388 (4.5)	598	273

*Results from the Bing API search after automatically filtering out duplicate results

+ Results from the manual Google Patent searches without duplicate filtering

++Results from the manual Google Patent searches after manually filtering out duplicate results

An extra manual check of the stratified sample of 560 Bing API results from all sixteen fields (35 results per field) gave an overall 98.4% precision (551 correct citation matches) for the automatic searches, identifying citations either in the non-patent reference sections (91.6% or 513) or in the main texts (6.8% or 38) of the full-text patents. From the 1.6% (9 of 560) false matches, four results were from patent citations with the same title and author (inventor). For instance, the query *Agrawal "Order preserving encryption for numeric data" 2004 site:google.com/patents* captured a patent citation with a partially identical title and with the same author (Patent number: US20050147240, *System and method for order preserving encryption for numeric data*, Inventor: Rakesh Agrawal). Other false matches were related to retrieving titles and author names in different parts of the patents (e.g., "*The invention relates to minimally invasive cardiac surgery." for the query Mack "Minimally invasive cardiac surgery*" 1996 in patent US8672998). Overall, however, the automatic method seems to give a high accuracy for the patent citation searches.

Google Patent Citation Counts

In all fields (Table 3) the vast majority of articles had no patent citations in the Bing results (and hence also in manual Google Patent searches; see above). Biomedical Engineering (10%), Biotechnology (9%), and Pharmacology & Pharmaceutics (7%) had the highest proportions of Scopus articles with at least one patent citation, suggesting that these fields either have a particularly direct commercial value or more of a patenting culture than others. These three field proportions at least triple those of Mechanical

Engineering (1.9%) and Energy Engineering (2.2%), showing that there are substantial disciplinary differences in the proportion of academic articles that are cited in patents.

The Relationship between Google Patents and Scopus Citations

There are statistically significant positive low correlations between Scopus and Google Patent citations from the Bing API searches in all fields (Table 3). Spearman correlations were used because citation data is typically skewed. Data for correlation tests should be separated by field and year (Fairclough & Thelwall, in press) for optimal accuracy, so see also Table 4 for more accurate results. The correlations are highest in Biomedical Engineering (.361), Biochemistry & Molecular Biology (.267), Biotechnology (.258), and Pharmacology & Pharmaceutics (.237) and lowest in Industrial & Manufacturing Engineering (.053), suggesting disciplinary differences in the relationship between patent and Scopus citations. The three high correlation subject areas had the highest proportions of articles with patent citations, suggesting that in the most patentable fields, traditional citations reflect commercial value most closely. Nevertheless, the low overall correlations suggest that the scientific research impact of academic articles is always only loosely related to their commercial impact, at least as reflected in patents, probably because applicants or patent examiners may tend to cite the most relevant sources such as other patents to justify their novelty contributions.

Table 3. Google Patent citations via the Bing API searches and Scopus citations for articles published every second year 1996-2012 (n=322,192 overall).

Scopus category	Articles (1996-2012)	Google Patent citations (% with citation) median (mean) max.	Scopus Citations (% with citation) median (mean) max.	Spearman correlation
Biochemistry & Molecular Biology	20,127	3,174, (**5.4%**) 0(0.16) 51	510,749, (80.4%) 8(25.4) 4,017	0.267**
Biomedical Engineering	20,076	8,846, (**10.1%**) 0(0.44) 149	449,581, (71.6%) 5(22.39) 3,429	0.361**
Biotechnology	20,169	5,488, (**9.2%**) 0(0.27) 58	469,916, (92%) 12(23.3) 2,381	0.258**
Chemical Engineering	20,170	1,451, (**2.9%**) 0(0.07) 39	306,995, (87.6%) 7(15.22) 964	0.135**
Computer Science	20,087	5,663, (**5.9%**) 0(0.28) 95	127,594, (58.2%) 1(6.35) 2,296	0.233**
Control & Systems Engineering	20,046	2,992, (**3.9%**) 0(0.15) 115	231,749, (58.1%) 1(11.56) 7,376	0.188**
Electrical & Electronic Engineering	20,174	4,153, (**5.6%**) 0(0.21) 63	383,897, (80%) 5(19.03) 4,532	0.199**
Energy Engineering	20,172	1,170, (**2.2%**) 0(0.06) 22	218,937, (60.8%) 2(10.9) 1,286	0.132**
Environmental Engineering	20,157	1,266, (**2.7%**) 0(0.06) 27	499,299, (90.8%) 11(24.77) 3,506	0.154**
Food Science	20,153	2,305, (**5.5%**) 0(0.11) 26	446,101, (94.1%) 12(22.14) 1,157	0.205**
Industrial & Manufacturing Engineering	20,192	1,504, (**2.5%**) 0(0.07) 67	147,145, (57.9%) 1(7.29) 815	0.053**
Mechanical Engineering	20,176	1,126, (**1.9%**)	277,360, (82.5%)	0.134**

		0(0.06) 33	6(13.75) 900	
Pharmacology & Pharmaceutics	20,195	3,777, (**6.8%**) 0(0.19) 82	438,882, (89.8%) 11(21.73) 1,408	0.237**
Physics Instruments & Instrumentation	20,062	1,848, (**2.8%**) 0(0.09) 87	185,033, (66.5%) 3(9.2) 688	0.160**
Polymer Science	20,124	2,552, (**4.6%**) 0(0.13) 41	329,639, (90.3%) 8(16.4) 1,181	0.195**
Surgery	20,112	2,010, (**2.9%**) 0(0.1) 45	385,463, (90.1%) 9(19.2) 1,171	0.158**

** Significant at the p = 0.01 level.

The weak but significant positive correlations between patents and Scopus citations across all sixteen science and engineering fields analysed suggests that academic papers are more likely to be commercially valuable if they are more highly cited (see also Tijssen, Buter, & van Leeuwen, 2000). The very weak relationship shows that academic citation counts would be a very poor substitute for patent citation counts, however, confirming that it is important to count the latter directly. The reason for the overall weak relationship might be that patents are less numerous and contain less citations than do journal articles, resulting in low patent citation counts and few papers with any patent citations at all. Unlike academics, inventors presumably also tend to be mainly interested in the most applied scientific research rather than purely theoretical contributions (Schmoch, 1993; Meyer, 2000a). In contrast, papers with high patent citation counts tend to be highly cited by academic publications (Meyer, Debackere, & Glänzel, 2010). Hence, it is useful to also check whether papers with high patent citation counts also tend to receive academic citations (see examples for the top patent citations across field, http://figshare.com/articles/GooglePatents_Appen_pdf/1418234). One example is the article "*Lind, R. et al. (1998). The network vehicle-a glimpse into the future of mobile multi-media. SAE Technical Papers",* which had no Scopus citations (as of March 2015) but had been cited in 67 patents. This may reflect the industry-focused nature of the *SAE Technical Papers* journal, but this does not undermine the contribution of the cited paper. Nevertheless, some patent citations do not reflect the technological contribution of the cited research (Jaffe, Trajtenberg, & Fogarty, 2000) and so qualitative evidence would be needed to confirm that an individual highly cited article had made a genuine commercial contribution.

Patent Citations by Year

Citations take time to accrue, whether from patents or academic articles, and so the publication year is important. There are higher Spearman correlations between Scopus and Google Patent citations for longer time periods in most fields (Table 4), presumably because the additional data makes the statistic more powerful. In Biomedical Engineering, for instance, 0.2% of the articles sampled from 2012 had one or more patent citations (r=.045) compared with 19.5% of the articles sampled from 1996 (r=.483). One reason for the large increases could be that it takes a long time for patents to be processed and granted – over two years for the USPTO (http://www.uspto.gov/learning-and-resources/general-faqs) – and industrial inventors may also be slower to patent than researchers are to publish, or may be less up-to-date with the academic literature if they are not publishing scientists.

From the results, long time periods are needed for assessing the commercial or technical value of academic publications based on patent citations (see also: Breschi et al., 2006). For the fields analysed, five years would be an absolute minimum since the figures are very low for 2012, and even fifteen years would give substantially more results than ten years. The length of these windows seem to rule out the use of patent citation analysis for current research but would nevertheless still be useful for longer term evaluations, such as that of the UK Research Excellence Framework, for which researchers must demonstrate societal impact for research that is up to 20 years old (REF, 2011).

Table 4. Articles with at least one Google Patent citation (Spearman correlation between Scopus and Google patent citations) by year (n=322,192 overall).

Fields	1996	1998	2000	2002	2004	2006	2008	2010	2012
Biochemistry & Molecular Biology	4.4% **(.297)**	6.1% **(.318)**	9.0% **(.328)**	7.7% **(.283)**	7.3% **(.262)**	6.2% **(.260)**	5.1% **(.245)**	2.5% **(.147)**	0.5% **(.071)**
Biomedical Engineering	19.5% **(.483)**	16.9% **(.411)**	13.3% **(.399)**	13.9% **(.337)**	13% **(.330)**	8.4% **(.237)**	4.3% **(.221)**	1.8% **(.163)**	0.2% (.045)
Biotechnology	13.3% **(.273)**	14.1% **(.267)**	13.3% **(.249)**	12.3% **(.260)**	10.5% **(.229**)	8.2% **(.224)**	6.0% **(.189)**	4.3% **(.168)**	0.8% **(.096)**
Chemical Engineering	5.0% **(.166)**	4.4% **(.123)**	3.8% **(.165)**	3.7% **(.159)**	3.1% **(.141)**	2.4% **(.137)**	1.9% **(.100)**	1.4% **(.094)**	0.4% **(.082)**
Computer Science	15.7% **(.293)**	8.6% **(.240)**	10.3% **(.325)**	7.2% **(.253)**	6.8% **(.247)**	3.2% **(.146)**	1.2% **(.113)**	0.4% **(.080)**	0.1% (.052)
Control & Systems Engineering	7.2% **(.226)**	7.7% **(.210)**	6.9% **(.194)**	5.1% **(.178)**	4.7% **(.175)**	2.1% **(.084)**	0.7% **(.057)**	0.3% **(0.061)**	0.04% (-.014)
Electrical & Electronic Engineering	10.2% **(.205)**	10.1% **(.247)**	9.5% **(.296)**	7.4% **(.230)**	6.2% **(.220)**	4.0% **(.150)**	1.9% **(.120)**	0.6% **(.057)**	0.1% (.005)
Energy Engineering	3.6% **(.142)**	3.1% **(.137)**	3.1% **(.152)**	3.0% **(.139)**	2.9% **(.158)**	1.6% **(.095)**	1.7% **(.090)**	0.5% (.016)	0.0% (-.015)
Environmental Engineering	5.0% **(.204)**	4.1% **(.159)**	3.6% **(.154)**	3.4% **(.187)**	2.2% **(.143)**	2.5% **(.136)**	2.3% **(.111)**	1.1% **(.105)**	0.4% (.014)
Food Science	8.8% **(.198)**	8.6% **(.239)**	8.5% **(.240)**	8.3% **(.198)**	5.9% **(.164)**	4.7% **(.153)**	3.1% **(.132)**	1.3% **(.103)**	0.3% (.050)
Industrial & Manufacturing Eng.	4.9% **(.096)**	4.4% **(.082)**	4.3% **(.076)**	3.4% **(.091)**	2.2% **(.064)**	1.6% **(.080)**	1.2% (.013)	0.5% (.040)	0.2% (-.015)
Mechanical Engineering	3.2% **(.130)**	3.2% **(.137)**	3.5% **(.190)**	2.3% **(.164)**	2.1% **(.114)**	1.4% **(.103)**	1.3% **(.136)**	0.3% **(.055)**	0.1% (.009)
Pharmacology & Pharmaceutics	8.8% **(.227)**	9.7% **(.239)**	9.9% **(.256)**	10.1% **(.235)**	8.3% **(.214)**	6.6% **(.245)**	4.6% **(.182)**	2.6% **(.134)**	0.7% **(.080)**
Physics Instruments	4.5% **(.165)**	6.2% **(.180)**	5.1% **(.224)**	3.8% **(.176)**	2.6% **(.154)**	1.5% *(.046)*	1.3% **(.111)**	0.5% (.027)	0.04% (-.017)
Polymer Science	6.9% **(.217)**	7.4% **(.192)**	6.7% **(.221)**	6.3% **(.224)**	5.4% **(.204)**	4.4% **(.186)**	2.9% **(.159)**	1.4% **(.083)**	0.2% (.021)
Surgery	4.9% **(.156)**	4.4% **(.170)**	4.3% **(.168)**	4.0% **(.150)**	3.3% **(.159)**	3.1% **(.153)**	1.7% **(.087)**	0.6% **(.058)**	0.1% (.011)

Bold figures: Correlations significant at the p = 0.01 level. Italic figures: Correlations significant at the p = 0.05 level. Normal figures: no statistically significant correlations between variables.

Results by Patent Office

The URLs of the Google Patent search results were used to assess the share of the citations that originated from the *US* (United States Patent and Trademark Office), *WO* (for World Intellectual Property Organization-WIPO), *EP* (for European Patent Office), *CA* (Canadian Intellectual Property Office), *CN* (State Intellectual Property Office of China) and *DE* (German Patent and Trade Mark Office). Over two thirds of the Google Patents citations in all fields were from US patents (Figure 1). This probably reflects the more extensive use of citations in USPTO patents than in the others covered as well and the English language selection criterion for the articles analysed. The relatively large size of the USPTO database is another important factor: since 1996 it has had at least three times as many applications per year as the other indexed offices except for China, which broke this pattern in about 2008 and grew rapidly to overtake the US in about 2010 (WIPO, 2014, p14). Biochemistry & Molecular Biology (69%) and Food Science (69%), had the lowest proportion of citations from US patents and Computer Science (98%) had the highest, probably reflecting (but exaggerating) the relative international dominance of the US in these fields. The duplicate removal process eliminated 3% of the citations, irrespective of patent office (see methods), and this may have had a minor impact on the proportions.

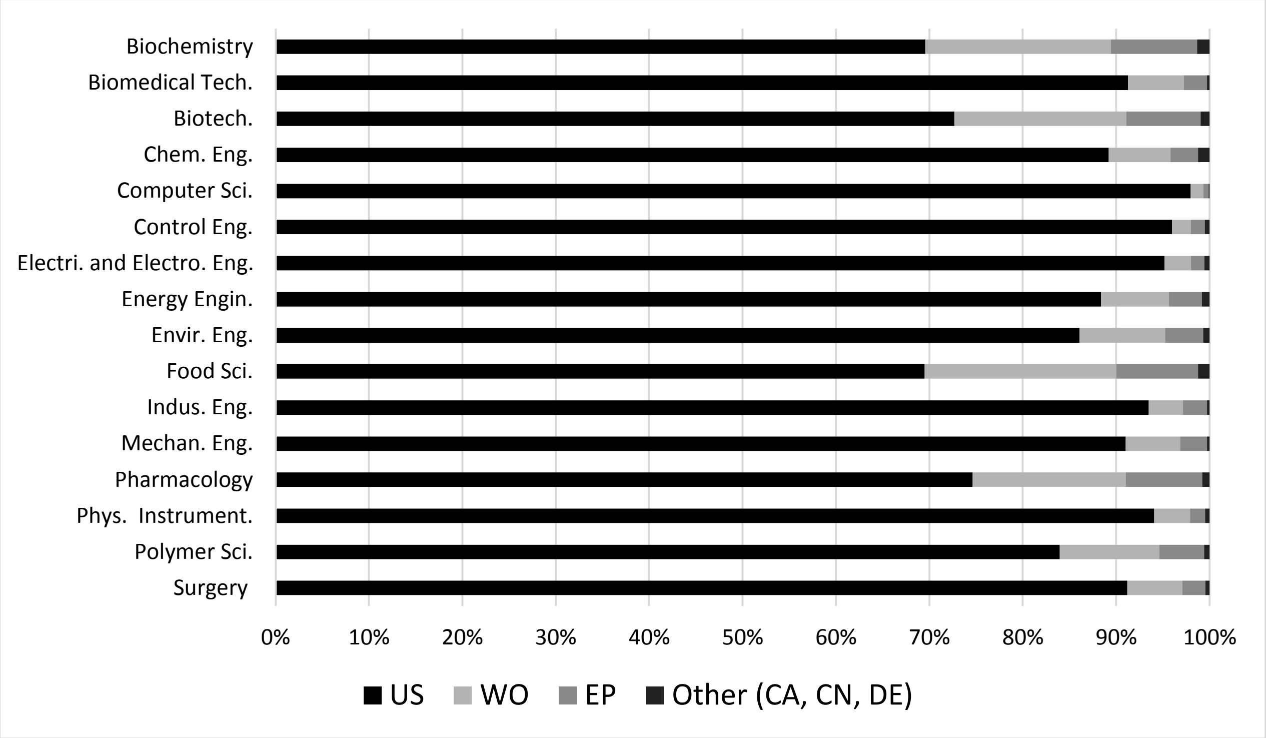

Figure 1. The patent office origins of the patent citations to the 322,192 Scopus articles investigated (n=322,192 overall).

Patent-Cited Topics and Scopus-Cited Topics

A simple heuristic was used to detect whether the topic areas that tended to attract patent citations were the same as those that tended to attract Scopus citations. For each subject area, all terms were extracted from the titles of the articles and converted to singular form, if plural. For each term, the number of times that it occurred in patent-cited articles was calculated and compared to the number of times it occurred in the remaining articles, using a difference in proportions test. The ten terms that occurred the most disproportionately often in patent-cited articles were then selected as potential indicators of important patentable topics within the field. To detect whether these terms were also associated with important Scopus citable topics, ranked lists were constructed in the same way for Scopus citations. Because patent citations are rarer than Scopus citations, a higher threshold should be used for Scopus citations, such as 10 or 100, but there is no obvious number to use. Since the objective was to identify topics that tended not to attract Scopus citations, a conservative approach was taken and three ranked lists were prepared: for 1 or more citations; for 10 or more citations; and for 100 or more citations. Also as a conservative approach, top 30 lists were prepared for each Scopus category rather than top 10 lists. Each top 10 patent-cited term was counted as a patent-specific topic if it did not occur in any of the three top 30 Scopus lists. Table 5 illustrates the process for one category, showing, for example, that the term *delivery* associates with patent citations and Scopus citations but the term *prodrug* only associates with patent citations. This approach is a simplification and is likely to be more effective for more frequent terms because, other factors being equal, they are more likely to occur with one of the top 30 Scopus lists if they are more frequent overall.

Table 5. The 10 most disproportionately used terms in the 1,369 patent cited articles out of the 20,195 articles in Pharmacology& Pharmaceutics.

Top 10 patent-cited term	Patent-cited articles (n=1369)	Non-patent cited articles (n=18826)	Patent-cited diff. in proportions test Z	Scopus 1+ citations top 30 rank*	Scopus 10+ citations top 30 rank*	Scopus 100+ citations top 30 rank*
delivery	6.1%(83)	2.3%(438)	8.4	-	-	9
prodrug	1.5%(21)	0.3%(51)	7.6	-	-	-
human	11.5%(157)	6.7%(1257)	6.7	5	2	20
inhibitor	6.4%(87)	3.0%(574)	6.6	13	21	-
encapsulated	0.7%(9)	0.1%(15)	6	-	-	-
nanoparticle	2.1%(29)	0.7%(127)	5.9	-	24	8
receptor	11.3%(155)	7.1%(1339)	5.7	1	1	7
novel	5.6%(76)	2.9%(545)	5.5	-	-	-
stabilized	0.4%(5)	0.0%(5)	5.4	-	-	-
polymeric	0.9%(12)	0.2%(32)	5.4	-	-	-

*Rank based on difference in proportions test Z-values for occurrence in the set of articles with at least *n* WoS citations

In each subject area there are terms that occur disproportionately often in patent-cited articles but not disproportionately often in more highly Scopus-cited articles (Table 6). The results suggest that there are topics, approaches or objects of study that are of particular commercial or applied relevance (or are more patentable) but are less cited. In addition, the results also suggest that there is a bigger difference between patentable and (academic) citable topics in some research areas (e.g., Electrical & Electronic Engineering) than others (e.g., Biomedical Engineering).

Table 6. The 10 most disproportionately used terms in the patent cited articles split by whether they are also in any one of three top 30 lists for academic citations.

Subject	Top ten patent-cited but not top thirty highly cited (x3)	Top ten patent-cited and top thirty highly cited (x3)
Biochemistry & Molecular Biology	photodynamic capsicum PCR angiogenesis pluripotent	Arabidopsis gene plant stem for
Biomedical Engineering	polyurethane cone-beam	hydrogel polymer tissue biodegradable calcium cement bone phosphate
Biotechnology	antibody recombinant libraries display antibodies hamster delivery	cell engineering novel
Chemical Engineering	biomaterial phospholipid PSA sieve lactate	CO2

	review negative complete colour	
Computer Science	wearable interface email WWW	rendering graphic speech access Java subdivision
Control and Systems Engineering	queries indexing video surgical micromachined reflectance appearance surgery	segmentation handwriting
Electrical & Electronic Engineering	MPEG-2 multi-resolution database power-efficient microprocessor constellation digital high-speed end-to-end	Video
Energy Engineering	stimulation insertion lead-acid microbial camera array	electrolyte lithium cell electrode
Environmental Engineering	switchgrass anaerobically digested hollow-fiber capture swing separation ethanol pretreatment	Membrane
Food Science	eicosapentaenoic docosahexaenoic administered phytosterol resveratrol orange higher	acid human cancer
Industrial & Manufacturing Eng.	diesel charge engine homogeneous ionization orbital power catalyst diode	alkali-activated
Mechanical Engineering	micromachined accelerometer polysilicon microactuator gyroscope resonator	silicon MEMS capacitive sensor
Pharmacology & Pharmaceutics	prodrug encapsulated novel stabilized polymeric	delivery human inhibitor nanoparticle receptor
Physics Instruments & Instrumentation	gyroscope polysilicon accelerometer SOI microsystem	micromachined bonding actuator MEMS cantilever
Polymer Science	resist flux extrusion benzoxazine oxetane polymeric	osmosis membrane gas protein
Surgery	society vacuum-assisted morphogenetic automated wound filter tissue matrix	engineering growth

Limitations

The results are subject to a number of limitations. Manual checks showed that the filtered Bing results missed 14% of the non-duplicate manual Google Patent matches (Table 2). The automatic patent searches therefore tend to reflect a subset of patent references covered by Google and are likely to be (small) underestimates of the total number of patent citations for articles. There may also be additional matches that Google does not return because it does not index the patent office or because its algorithm does not always return all results. Although Webometric Analyst allows up to 5,000 patent citation searches per month by having a free Windows Azure Marketplace key, for a larger patent citation searchers a fee-based subscription is required (see https://datamarket.azure.com/dataset/bing/search).

Scopus articles with one or two word titles were excluded from the searches due to the prevalence of false matches and non-articles documents with short titles. This method excluded some relevant research articles (e.g., Short-bowel syndrome). Hence, future applications may try alternative queries, perhaps adding extra bibliographic information to the queries, such as the last names of the three first authors or common journal name abbreviations. Moreover, sometimes article titles contain non-alphanumeric characters such as *"…HBr(B1Σ+) emission from the ultraviolet…"* or *"…analysis using lab VIEW® and charm works™"* which might undermine some searches. We excluded non-English Scopus articles from our dataset and this may undermine Google Patent citations from non-English patents, especially for the European, German Chinese patent offices.

The method to filter out duplicate matches from the initial Bing API results is another limitation because some different patents have the same titles and descriptions in Bing. For instance, the query *Olsson "Fermentation of lignocellulosic hydrolysates for ethanol production" 1996 site:google.com/patents* returned four results from three different patent offices with the same title (*Polypeptides having cellobiohydrolase activity and polynucleotides encoding same*) and description and so the automatic filtering method only retained one of them (US8759023B2; US8771994B1; WO2010141325A1; EP2668270A1). Manual checks showed that these patents were for different chemicals despite having the same titles, very similar wordings throughout, and citing the above article in the same context. For example, US8759023B2 is for, in part, "a polypeptide having at least 90% sequence identity to the sequence of amino acids 26 to 532 of SEQ ID NO: 4" and US8771994B1 is for, in part, "an isolated polypeptide, which has at least 95% sequence identity to the sequence of amino acids 21 to 446 of SEQ ID NO: 2". The patents were presumably copied from the first one with words changed only when necessary to reflect the new invention. Future research might be able to construct a method to avoid removing such matches, although there were only 3% duplicate citations removed and so it would have little impact on the overall results.

The results might be affected by the choice of Scopus as the data source. To check for this, the Google Patent citations were compared to both WoS articles with Scopus articles published in 2004 across the same subject categories (Table 7). A random sample of WoS articles were taken in each selected field and the same automatic method was used to locate citations in Google Patents indexed by Bing and there were some differences between the samples from the two databases. The results suggest that WoS articles are more likely to receive patent citations than are Scopus articles, perhaps because WoS is more selective than Scopus. The correlations do not vary much according to the database used, however.

Table 7. Secondary analyses of Google Patent citations to WoS-indexed and Scopus-indexed articles in 2004.

Discipline (Scopus categories)	Sampled articles published in 2004 from WoS (Scopus)	% of articles with one or more Google Patent citations from WoS (Scopus)*	Correlations between Google Patents and journal citations WoS (Scopus)**
Biochemistry & Molecular Biology	2,236 (2,223)	**10.7%** (7.3%)	0.241 (0.262)
Biomedical Engineering	2,241 (2,233)	**15.4%** (13%)	0.309 (0.330)
Biotechnology	2,231 (2,241)	**11.7%** (10.5%)	0.216 (0.229)
Chemical Engineering	2,229 (2,236)	**4.8%** (3.1%)	**0.170** (0.135)
Computer Science	2,232 (2,240)	**8.6%** (6.8%)	**0.268** (0.241)
Control and Systems Engineering	1,458 (2,190)	4% (4.7%)	**0.191** (0.165)
Electrical & Electronic Engineering	2,235 (2,245)	4.6% (6.2%)	0.162 (0.220)

Energy Engineering	2,230 (2,234)	**4.7%** (2.9%)	**0.170** (0.160)
Environmental Engineering	2,235 (2,233)	**3%** (2.2%)	**0.139** (0.136)
Food Science	2,239 (2,235)	**6.3%** (5.9%)	**0.172** (0.157)
Industrial & Manufacturing Engineering	2,239 (2,271)	1.9% (2.2%)	**0.099** (0.064)
Mechanical Engineering	2,237 (2,233)	**1.6%** (2.1%)	**0.135** (0.114)
Pharmacology & Pharmaceutics	2,243 (2,233)	**9.3%** (8.3%)	0.089 (0.203)
Physics Instruments & Instrumentation	1,701 (2,225)	**4%** (2.6%)	**0.191** (0.154)
Polymer Science	2,237 (2,241)	**6.2%** (5.4%)	0.204 (0.206)
Surgery	2,241 (2,232)	**3.4%** (3.3%)	0.137 (0.154)

*Higher values from WoS compared with Scopus are highlighted in bold.

**All Spearman correlations are significant at the p = 0.01 level.

Conclusions

In answer to the first research question, the new method automates the process of identifying patent citations to academic papers with a high degree of accuracy (98%) and coverage. Although 13% of the manual Google patent results were not found, this is a relatively small percentage. Conventional citation databases are unable to track the commercial impacts of articles and this is especially important in some engineering fields that often produce academic research that directly leads to valuable inventions. This gap can be partly filled by patent citations, and this article has introduced a practical method to semi-automatically extract patent citations from reasonably comprehensive and complete patent databases. It is free for small collections of articles but for large collections, the additional queries necessary must be purchased from Microsoft. Some additional manual searches and checking may be needed for articles with short or complex titles.

In answer to the second research question, the low but statistically significant correlations between the Google Patents citations and Scopus citations to academic articles across all fields suggest that patent citations and traditional citations predominantly reflect different types of research contribution. It is therefore important to gather patent citations when evaluating research in order to avoid disadvantaging research with value that is not fully reflected within academia. The same is true for broad research topics (RQ5) since in all fields there are some that attract patent citations but not many traditional citations. Hence, funders should be careful if withdrawing funding from low cited research areas because some may attract patent citations instead.

In answer to the third research question, there were clear disciplinary differences in the proportions of articles with at least one patent citation. This supports the previous finding that patents in fast growing subject areas, such as genetics and biotechnology, tend to cite more scientific literature (Narin & Noma, 1985; Collins & Wyatt, 1988; Lo, 2010; Tijssen, 2001). The fields chosen for the study were all in commercially relevant areas and so the disciplinary differences are likely to be much larger across a wider range of fields. Hence, it would clearly not be fair to compare patent citations between different similar fields for evaluation purposes. The results are also highly dependent on time, and an absolute minimum of 5 years is needed to get substantial patent citation results, and much better results would be obtained for a fifteen year window. Policymakers should therefore not expect academic research to translate into commercial value in the short to medium term and should allow long time periods for commercial transfer evaluations (as in the UK REF impact case studies).

The patent results from Google originate predominantly from the USPTO (RQ4) and so the method here should not be used in contexts where substantial numbers of relevant patents are expected to have been filed exclusively elsewhere. One reason for the large number of citations from the US patents could be that Google has greater coverage of the US patents than of EPO and WIPO patents. However, the extent to which non-patents sources are cited in patents may also be influenced by the different regulations and examination procedures in patent offices (see Michel & Bettels, 2001; Callaert et al, 2006). Biochemical Engineering articles were most likely to attract patent citations (10%) and Mechanical Engineering articles were the least likely (2%), despite both being applied types of research that could reasonably be expected to generate patentable inventions. Moreover, the most patented areas were not the same as the most Scopus-cited areas (RQ5): for example, Mechanical Engineering articles (83%) were more likely to attract Scopus citations than were Biomedical Engineering articles (72%). The method is thus likely to be irrelevant in fields (including some applied research areas) that tend not to patent inventions. Other than these limitations, it seems that the new method is a valuable new tool for patent analysis to help evaluators assess the long term commercial impacts of academic research in some fields.

References

Alcácer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. Review of Economics and Statistics, 88(4), 774-779.

Alcácer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. Research Policy, 38(2), 415-427.

Amara, N., & Landry, R. (2012). Counting citations in the field of business and management: Why use Google Scholar rather than the Web of Science. Scientometrics, 93(3), 553-581.

Bar-Ilan, J. (2008). Which h-index? - A comparison of WoS, Scopus and Google Scholar. Scientometrics, 74(2), 257-271.

Beel, J., & Gipp, B. (2010). On the robustness of Google Scholar against spam. Paper presented at the HT'10 - Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, 297-298.

Breschi, S., Trasconi, G., Catalini, C., Novella, L., Guatta, P., & Johnson, H. (2006). Highly cited patents, highly cited publications, and research networks. Investment in Research and Links with other Policies, European Commission. http://ec.europa.eu/invest-in-research/pdf/download_en/final_report_hcp.pdf

Brusoni, S., Criscuolo, P., & Geuna, A. (2005). The knowledge bases of the world's largest biopharmaceuticals groups: what do the patent citations to non-patent literature reveal? Economics of Innovation and New Technology, 14(5), 395–415.

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.

Callaert, J., Pellens, M., & Van Looy, B. (2014). Sources of inspiration? Making sense of scientific references in patents. Scientometrics, 98(3), 1617-1629.

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.

Cohen, W. M. (2010). Fifty years of empirical studies of innovative activity and performance. In B. Hall and N. Rosenberg (eds.). The Economics of Innovation (Amsterdam: Elsevier), 129–213.

Collins, P., & Wyatt, S. (1988). Citations in patents to the basic research literature. Research Policy, 17(2), 65-74.

Courtial, J. P., Callon, M., & Sigogneau, A. (1993). The use of patent titles for identifying the topics of invention and forecasting trends. Scientometrics, 26(2), 231-242.

Cronin, B. (2001). Bibliometrics and beyond: Some thoughts on web-based citation analysis. Journal of Information Science, 27(1), 1-7.

De Groote, S. L., & Raszewski, R. (2012). Coverage of Google scholar, Scopus, and Web of Science: A case study of the h-index in nursing. Nursing Outlook, 60(6), 391-400.

de Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: A longitudinal study. Scientometrics, 98(2), 1547-1565.

Fairclough, R. & Thelwall, M. (in press). The influence of time and discipline on the magnitude of correlations between citation counts and quality scores. Journal of Informetrics.

Guan, J. C., & He, Y. (2007). Patent-bibliometric analysis on the Chinese science—Technology linkages. Scientometrics, 72(3), 403–425.

Han, H., Xu, S., Zhu, L., Qiao, X., Gui, J., & Zhang, Z. (2014). Mining technical topic networks from Chinese patents. Paper presented at the CEUR Workshop Proceedings, 1292.

Harzing, A. , & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. Ethics in Science and Environmental Politics, 8(1), 61-73.

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.

Huang, M., Huang, W., & Chen, D. (2014). Technological impact factor: An indicator to measure the impact of academic publications on practical innovation. Journal of Informetrics, 8(1), 241-251.

Hung, W. C. (2012). Measuring the use of public research in firm R&D in the hsinchu science park. Scientometrics, 92(1), 63-73.

Jacsó, P. (2011). Google Scholar duped and deduped - the aura of "robometrics". Online Information Review, 35(1), 154-160.

Jaffe, A., Trajtenberg, M., & Fogarty, M. (2000). The meaning of patent citations: Report on the NBER/Case-Western Reserve Survey of Patentees. NBER Working Papers No. 7631. http://www.nber.org/papers/w7631.pdf

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. PLoS ONE, 9(5): e93949. doi:10.1371/journal.pone.0093949.

Kousha, K. & Thelwall, M. (in press). An automatic method for assessing the teaching impact of books from online academic syllabi. Journal of the Association for Information Science and Technology.

Kousha, K., & Thelwall, M. (2007a). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. Journal of the American Society for Information Science and Technology, 58(7), 1055-1065.

Kousha, K., & Thelwall, M. (2007b). How is science cited on the web? A classification of Google unique web citations. Journal of the American Society for Information Science and Technology, 58(11), 1631-1644.

Kousha, K., & Thelwall, M. (2008). Assessing the impact of disciplinary research on teaching: An automatic analysis of online syllabuses. Journal of the American Society for Information Science and Technology, 59(13), 2060-2069.

Kousha, K., & Thelwall, M. (2009). Google Book Search: Citation analysis for social science and the humanities. Journal of the American Society for Information Science and Technology, 60(8), 1537-1549.

Kousha, K., & Thelwall, M. (2014). An automatic method for extracting citations from Google Books. Journal of the Association for Information Science and Technology, 66(2), 309–320.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. Journal of the American Society for Information Science and Technology, 62(11), 2147-2164.

Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. JAMA - Journal of the American Medical Association, 302(10), 1092-1096.

Lawson, M., Kemp, N., Lynch, M. F., & Chowdhury, G. G. (1996). Automatic extraction of citations from the text of English-language patents - an example of template mining. Journal of Information Science, 22(6), 423-436.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. Technovation, 29(6-7), 481-497.

Lemley, M. A., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. Review of Economics and Statistics, 94(3), 817-827.

Leydesdorff, L. (2004). The university-industry knowledge relationship: Analyzing patents and the science base of technologies. Journal of the American Society for Information Science and Technology, 55(11), 991-1001.

Li, R., Chambers, T., Ding, Y., Zhang, G., & Meng, L. (2014). Patent citation analysis: Calculating science linkage based on citing motivation. Journal of the Association for Information Science and Technology, 65(5), 1007–1017.

Liaw, Y., Chan, T., Fan, C., & Chiang, C. (2014). Can the technological impact of academic journals be evaluated? The practice of non-patent reference (NPR) analysis. Scientometrics. DOI 10.1007/s11192-014-1337-0

Lo, S. (2010). Scientific linkage of science research and technology development: A case of genetic engineering research. Scientometrics, 82(1), 109-120.

Lopez, P. (2010). Automatic extraction and resolution of bibliographical references in patent documents. In H. Cunningham, A. Hanbury & S. Rüger, eds., Advances in Multidisciplinary Retrieval, vol. 6107 of Lecture Notes in Computer Science, 120−135, Springer Berlin / Heidelberg.

López-Cózar, E. D., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. Journal of the Association for Information Science and Technology, 65(3), 446-454.

Ma, Y., Sun, L., Wang, Z., & Yang, X. (2010). Computer aided system of screening and sorting in data processing for non-patent literature. Paper presented at the ICCASM 2010 - 2010 International Conference on Computer Application and System Modeling, Proceedings, 3 V3609-V3611.

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. Journal of the American Society for Information Science and Technology, 58(13), 2105-2125.

Meyer, M. (2000a). What is special about patent citations? Differences between scientific and patent citations. Scientometrics, 49(1), 93-123.

Meyer, M. (2000b). Does science push technology? Patents citing scientific literature. Research Policy, 29(3), 409-434.

Meyer, M. (2003). Academic patents as an indicator of useful research? A new approach to measure academic inventiveness. Research Evaluation, 12(1 SPEC ISS.), 17-27.

Meyer, M., Debackere, K., & Glänzel, W. (2010). Can applied science be 'good science'? Exploring the relationship between patent citations and citation impact in nanoscience. Scientometrics, 85(2), 527-539.

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.

Nagaoka, S. (2007). Assessing the R&D management of a firm in terms of speed and science linkage: Evidence from the US patents. Journal of Economics and Management Strategy, 16(1), 129-156.

Nanba, H., Anzen, N., & Okumura, M. (2008). Automatic extraction of citation information in Japanese patent applications. International Journal on Digital Libraries, 9(2), 151-161.

Narin, F., & Noma, E. (1985). Is technology becoming science? Scientometrics, 7(3-6), 369-381.

Narin, F., & Olivastro, D. (1992). Status report: Linkage between technology and science. Research Policy, 21(3), 237-249.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. Research Policy, 26(3), 317-330.

Oppenheim, C. (2000). Do patent citations count? In: B. Cronin & H B. Atkins (Eds.), The web of knowledge: A festschrift in honor of Eugene Garfield (pp. 405-432). Metford, NJ. Information Today Inc ASS Monograph Series.

REF (2011). Assessment framework and guidance on submissions. http://www.ref.ac.uk/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf

Roach, M., & Cohen, W. M. (2013). Lens or prism? Patent citations as a measure of knowledge flows from public research. Management Science, 59(2), 504-525.

Sampat, B. (2004) Examining patent examination: an analysis of examiner and applicant generated prior art. National Bureau of Economic Research Summer Institute Presentation. http://www.stiy.com/MeasuringInnovation/Sampat.pdf

Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. Scientometrics, 26(1), 193-211.

Shirabe, M. (2014). Identifying SCI covered publications within non-patent references in U.S. utility patents. Scientometrics, DOI 10.1007/s11192-014-1293-8

Subramanian, A. M., & Soh, P. (2010). An empirical examination of the science-technology relationship in the biotechnology industry. Journal of Engineering and Technology Management - JET-M, 27(3-4), 160-171.

Thelwall, M., & Kousha, K. (2008). Online presentations as a source of scientific impact? An analysis of PowerPoint files citing academic journals. Journal of the American Society for Information Science & Technology, 59(5), 805-815.

Thelwall, M., & Maflahi, N. (2015). Guideline references and academic citations as evidence of the clinical value of health research. Journal of the Association for Information Science and Technology. DOI: 10.1002/asi.23432

Tijssen, R. J. W. (2001). Global and domestic utilization of industrial relevant science: Patent citation analysis of science-technology interactions and knowledge flows. Research Policy, 30(1), 35-54.

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. Retrieved from www.scopus.com

Tseng, Y., Lin, C. & Lin, Y. (2007). Text mining techniques for patent analysis. Information Processing and Management, 43(5), 1216-1247.

Van Looy, B., Zimmermann, E., Veugelers, R., Verbeek, A., Mello, J., & Debackere, K. (2003). Do science-technology interactions pay off when developing technology? An exploratory investigation of 10 science-intensive technology domains. Scientometrics, 57(3), 355-367.

van Vianen, B. G., Moed, H. F., & van Raan, A. F. J. (1990). An exploration of the science base of recent technology. Research Policy, 19(1), 61-81.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: What is the difference? Journal of the American Society for Information Science and Technology, 54(14), 1313-1322.

Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. Journal of the American Society for Information Science and Technology, 56(10), 1075-1087.

Verbeek, A., Debackere, K., Luwel, M., Andries, P., Zimmermann, E., & Deleus, F. (2002). Linking science to technology: Using bibliographic references in patents to build linkage schemes. Scientometrics, 54(3), 399-420.

WIPO (2014). World Intellectual Property Indicators. http://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2014.pdf

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. Journal of High Technology Management Research, 15(1), 37-50.