Web citations in patents: Evidence of technological impact?[1]

Enrique Orduna-Malea*
EC3 Research Group, Universitat Politècnica de València (UPV), 46022 Valencia, Spain. E-mail: enorma@upv.es, Tel: +34 963879480

Mike Thelwall and Kayvan Kousha
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

Abstract
Patents sometimes cite web pages either as general background to the problem being addressed or to identify prior publications that will limit the scope of the patent granted. Counts of the number of patents citing an organisation's website may therefore provide an indicator of its technological capacity or relevance. This article introduces methods to extract URL citations from patents and evaluates the usefulness of counts of patent web citations as a technology indicator. An analysis of patents citing 200 US universities or 177 UK universities found computer science and engineering departments to be frequently cited, as well as research-related web pages, such as Wikipedia, YouTube or Internet Archive. Overall, however, patent URL citations seem to be frequent enough to be useful for ranking major US and the top few UK universities if popular hosted subdomains are filtered out, but the hit count estimates on the first search engine results page should not be relied upon for accuracy.

Keywords: Link analysis; Webometrics; Search engines; Google Patents; Universities; United States; United Kingdom.

Introduction

Citations in patents can potentially be used to assess connections between science and industry (Narin & Olivastro, 1992; Narin, Hamilton, & Olivastro, 1997). Although references in patents commonly cite other patents and non-patent sources such as journal articles, conference papers and books (Oppenheim, 2000), some patents have links to online sources (web citations). Thus, link analyses of digitised patents may reveal knowledge flows between academia and industry, and informational connections between inventors and online content providers. For instance, high patent citation counts may be used either by universities, departments and companies as evidence of their achievements with emerging technologies and innovations.

This motivation echoes previous attempts to use hyperlink counts as impact indicators in academic contexts (Cronin, 2001; Brin & Page, 1998; Ingwersen, 1998), including for university websites (Thelwall, 2004). Nevertheless, this idea for a new type of indicator must be assessed to see whether it is practical and gives meaningful figures.

The Google Patents website (Banks, 2006; Endres, 2007) is a logical source from which to extract URL patent citations. It indexes the full text of patents and patent applications from: United States Patent and Trademark Office (UPSTO), European Patent Office (EPO), World Intellectual Property Organization (WIPO), Deutches Patent und Markenamt (DPMA), Canadian Intellectual Property Office (CIPO), and

China's State Intellectual Property Office (SIPO). As of April 2016, Google Patents covers 73,461,560 (41,716,473 grants and 31,745,087 applications), with an important percentage of Japanese and Chinese and US patents (34.8%, 19.6%, and 19.9% respectively) (patents.google.com). Although Google Patents does not index all patent offices, its full-text search capability (Marley, 2014) makes it possible to search huge numbers of digitised patents including for URL citations. Furthermore, the fact of being a controlled environment prevents the existence of spam, a shortcoming of other types of link analysis (e.g., Orduna-Malea, 2013).

Some previous studies have assessed Google Patents for types of impact indicator. In particular counts of patents granted have been proposed as an innovation indicator for businesses (Moskovkin, Shigorina and Popov, 2012), and a semi-automatic method has been developed to extract Google Patent citations to academic articles as the commercial impact indicator (Kousha and Thelwall, in press). Nevertheless, no previous investigation has assessed URL citations in online patents as a source of technological indicators for academic institutions, or inventors' uses of hyperlinks to online resources.

Background

Link-based scholarly indicators

Many previous studies have investigated whether counts of the number of hyperlinks pointing to academic websites could be a useful indicator of some type of wider impact. Link counts significantly correlate with research productivity indicators for universities within a single country (Thelwall, 2001; Smith & Thelwall, 2002), although they may primarily reflect institutional visibility (Thelwall & Kousha, 2015). These correlations occur despite some hyperlinks being generated automatically for navigation, publicity, and other non-scholarly reasons (Wilkinson, Harries, Thelwall, & Price, 2003). Hyperlink counts have also been used to construct ranked lists of universities based upon their web presences (Aguillo, Ortega & Fernández, 2008).

Although it is no longer possible to exploit major search engines to identify hyperlinks, alternative methods have been proposed to identify citations to other websites, including URL citations, title mentions and linked title mentions using commercial search engines (Kousha & Thelwall, 2006; Thelwall & Sud, 2011; Ortega, Orduna-Malea & Aguillo, 2014; Sud & Thelwall, 2014). An URL citation is a mention of an URL in the text of a webpage, whether or not it is hyperlinked. Commercial search engines can be queried for such URL citations, allowing URL citations counts to be harvested via automated queries for large sets of websites.

An alternative to using web link variants is to use social media metrics such as commenting, downloading, and recommending, known as altmetrics (Priem & Hemminger, 2010; Thelwall, 2012). Some of these metrics include hyperlinks, for example within Tweets (Orduna-Malea, Torres-Salinas & Delgado López-Cózar, 2015; Vaughan, 2015), but most focus on impact indicators for individual articles or journals (Haustein & Siebenlist, 2011) rather than for entire organizations. The number of links from Twitter (Orduna-Malea, Torres-Salinas & Delgado López-Cózar, 2015) and Wikipedia (Orduna-Malea & Ontalba-Ruipérez, 2013) correlate with the inlinks received by sets of national and international universities, giving evidence of their value.

In order to be able to interpret the meaning of hyperlink and other indicators, positive significant correlations with established indicators are not enough (Seeber et al., 2012) and content analyses are also needed to get insights into why links are created (Bar-Ilan, 2004a; Thelwall, 2004). Link motivation studies have found many different

reasons why hyperlinks are created, mostly related to the main activities of the target organization (Smith, 1999; Kim, 2000; Thelwall, 2002; Park, 2002; Harries, Wilkinson, Price, Fairclough & Thelwall, 2004). For example, nearly all (90%) links between UK universities are created for scholarly-related activities (Wilkinson, Harries, Thelwall & Price, 2003), although there seem to be more general-purpose links between Israeli universities (Bar-Ilan, 2004a; 2005). Thus, whilst counts of hyperlinks to academic-related web sites associate with research quality, this does not imply a cause-and-effect relationship because most hyperlinks do not directly target research outputs.

Patent citations

Patents are technical documents that register their inventions and confer grantees the right to legally protect them for a limited period and within a specific geographic area (e.g., Europe, the USA). Patent applications can, but do not have to, include references to prior patents (known as *patent citations*) or other documents (*nonpatent citations*). Citations from patents to scientific articles, monographs, proceedings papers or working reports can be used to trace connections between academia and business, perhaps reflecting direct or indirect knowledge transfer (Narin & Noma, 1985; Narin, Hamilton & Olivastro, 1997; Tijssen, Buter, & van Leeuwen, 2000; Verbeek et al., 2002).

Citations in patents are not always created for scientific reasons. Inventors may cite non-patent documents either as general background to the problem being addressed or to identify prior publications that will limit the scope of the patent granted. Moreover, examiners can also add or remove references in order to clarify an invention's novelty (Van Looy et al, 2006; Alcácer, Gittelman & Sampat, 2009). This makes the analysis of citations in patents a distinctive bibliometric process with wider citing motivations (Meyer, 2000a; 2000b; 2003; Li et al, 2014), and substantial disciplinary differences (Kousha & Thelwall, in press). For this reason, the use of patent URL citations may perhaps have similar functions to those of general nonpatent citations. Furthermore, the connection established between the patent (and aggregates such as inventors, topics, and assignees) and the resources linked (and aggregates such as sources, authors, etc.) can stimulate patent citation research. The technical proof of concept of the process of extracting URLs from patent references has not been tested to date, however.

Research questions

This article introduces and tests novel methods to semi-automatically extract URL citations from patent documents on a large scale from the public domain contents provided by Google Patents in order to determine their degree of use by inventors and evaluate them as evidence of technological impact. Search engine hit count estimates are known to be unreliable for general search engines (Rousseau, 1999; Bar-Ilan, 2004b; Thelwall, 2008), and so it is important to evaluate them, as a convenient source of URL information, for the Google Patents search engine. The main research questions are the following:

- (RQ1) Can the Google Patents search interface be used to give accurate counts of URLs in Google patents?
- (RQ2) Are there enough URL citations to university websites in patents for use in impact indicators?
- (RQ3) Do the types of university web pages cited by patents associate with a useful type of impact?
- (RQ4) Are URL citations frequently used in patents by inventors?

Methods

The study was restricted to documents matching the following parameters in order to give a coherent scope and to avoid duplicate patents submitted to different offices. The US Patent office was chosen as a testbed because it seems to include the most patents, its patents frequently include non-patent references (Michel & Bettels, 2001) and its patents are not automatically translated.

- Patent Office: US Patent Office (USPTO)
- Filing status: issued document
- Filing date: January 1st, 2005 to December 31, 2014

The websites of UK and US universities were used to match URLs in patents. These countries are presumably the most cited in US patents, although Canada and Germany may also be frequently cited. Since web presence is important for this, the top 200 national universities listed in the most recent edition of the Ranking Web of World Universities (January 2015) for USA and UK were selected. Only institutions offering graduate degrees were considered. The final set included 200 US, and 177 UK universities.

For the first research question three different methods were devised to search for URLs within Google Patents, and their results compared with each other and a (much more time consuming) crawler-based method using the Google Patents sitemap complete list of patent pages (https://www.google.com/patents/sitemap/en/Sitemap.html). In theory, the crawler method should give comprehensive results because the entire database was crawled and all matching URLs were counted, whereas the SERP methods may give estimates due to the use of approximation algorithms. Nevertheless, the process of matching URLs in webpages is not straightforward because there may be scanning errors and incomplete URLs and so none of the methods are guaranteed to give accurate or comprehensive results.

- SERP100_1: A manual query (e.g., “harvard.edu”) in the Google Patents search interface (timespan: 2005 to 2014; 1 query per URL; 100 results per results page) for each of the selected UK and US universities, using the hit count estimate in the first search engine results page (SERP).
- SERP100_n: As above but using the hit count estimate in the final SERP.
- SERP10_n: As above but requesting 10 results per page.
- Crawler: Google Patents crawled by SocSciBot (http://socscibot.wlv.ac.uk) and URLs in these documents automatically extracted using Webometric Analyst (http://lexiurl.wlv.ac.uk, the *Google Patents: Extract SocSciBot crawl info* option in the *Services* menu) and matched to the domains of the selected US and UK universities.

The Google Custom API can also be used as an alternative method to collect hit estimate numbers from SERPs. However, search results are limited to 100 per day free (https://developers.google.com/custom-search/json-api/v1/overview).

Google Patents offers two different versions for each existing patent document: a PDF and a HTML version. Only the HTML versions were crawled in this study to avoid duplication.

The first three methods were conducted from 15 to 30 April 2015 and the crawler worked during May 2015. Since only the crawl gave lists of URLs, rather than counts of URLs, only the crawler data was used for analyses of individual URLs.

Results

RQ1: Counts of website URLs in Google Patent searches

The four methods give substantially different estimated total numbers of matches (Table 1). Since the crawler method should be reasonably accurate, it seems clear that SERP100_1 can give unrealistic overestimates, whereas both SERP100_n and SERP10_n give much more plausible figures. Thus, the estimated total number of matches in the first search results page seems to be too optimistic to be of practical use.

TABLE 1. The sum (median) of the number of patents containing URLs of the selected universities from each of the four methods.

University set	SERP100_1	SERP100_n	SERP10_n	Crawler
US (n=200)	649,671 (78)	19,935(78)	28,560 (72)	28,900 (64)
UK (n=177)	30,591 (0)	3,135 (0)	3,729 (0)	2,667 (0)

Despite the differences in results between methods, the extremely high correlations between them suggest that they are interchangeable in practice for the purpose of ranking universities, at least for the high profile universities considered here (Table 2). Thus, the level of apparent inflation in the first SERP seems to be consistent between websites. In fact, it only seems to apply to websites with many hits (similar to: Thelwall, 2008), which would not affect the rankings much. This is partly because the SERP100_1 and SERP100_n are identical for websites with under 100 hits and partly because the SERP100_1 appear to be unrealistically large in some cases when there are more than 100 results, as found by the crawler. The mean difference between the SERP100_1 and crawler data for the US is 3091 (median 18). There is little to choose between SERP100_n and SERP10_n: although the mean difference from the crawler data is larger for SERP100_n (57.5) than for SERP10_n (32.0), they both have difference of 11. The UK follows a similar pattern in terms of average differences with the crawler data (SERP100_1: 157.8; SERP100_n: 4.9; SERP10_n: 6.0; all medians are zero) except that SERP100_n is marginally better than SERP10_n. Hence, overall there is no real evidence about which of SERP100_n and SERP10_n is best overall.

TABLE 2. Spearman correlations between the different search methods. Top-right: triangle: UK universities; bottom right triangle: US universities.

Method	SERP100_1	SERP100_n	SERP10_n	Crawler
SERP100_1	-	**1.00	**1.00	**0.98
SERP100_n	**0.99	-	**1.00	**0.98
SERP10_n	**0.99	**1.00	-	**0.98
Crawler	**0.96	**0.97	**0.97	-

** Significant at alpha=0.001

RQ2: Total number of academic website URLs in USPTO patents

The distribution of patent URL citations according to the targeted university is highly skewed (see Tables 3, 4). For US universities, whilst five universities have URL citations in over 1000 patents from the crawler results, 123 have less than 100 patent URL citations and 2 have none. For the UK there are fewer patent URL citations overall, 168 have less than 100 patent URL citations and 99 have none. The much higher level of citing for the US is only partly due to the choice of the USPTO as the patent source because US universities also attract many more citations in EPO results (not shown). A possible reason for this is that many EPO citations may originate from

US-authored USPTO patents that are duplicated in EPO but retain the additional citing information that is common in the USPTO system. Hence, patent citations may only be common enough to help rank major US universities and top UK universities.

TABLE 3. US universities most cited in patents (crawler data).

US universities	SERP100_1	SERP100_n	SERP10_n	Crawler	THES*
Pennsylvania State University	64,700	434	784	**1,661**	23rd
Massachusetts Institute of Technology	65,200	433	787	**1,493**	1st
Stanford	48,700	387	780	**1,175**	2nd
Carnegie Mellon	35,200	294	633	**1,082**	10th
University of California Berkeley	40,000	347	698	**1,012**	6th
University of Maryland	16,300	236	477	**586**	31th
University of Illinois-Urbana Champaign	29,500	339	594	**561**	12th
University of Washington	22,200	399	572	**540**	17nd
Cornell	12,600	264	476	**530**	11th
Georgia Institute of Technology	20,400	236	433	**482**	7th

* 2014-2015 Times Higher Education World University Rankings for engineering and technology (United States).

TABLE 4. UK universities most cited in patents (crawler data).

UK universities	SERP100_1	SERP100_n	SERP10_n	Crawler	THES*
Cambridge	12,700	252	456	**303**	1st
Edinburgh	5,830	175	340	**283**	6th
Newcastle	1,160	189	255	**210**	--
Oxford	2,520	151	208	**162**	3nd
UCL	2,910	164	235	**138**	5th
Glasgow	1,430	121	147	**125**	--
Queen Mary, London	625	134	180	**122**	--
Surrey	501	105	100	**87**	--
Sussex	491	112	97	**81**	--
Aberdeen	806	114	105	**74**	--

* 2014-2015 Times Higher Education World University Rankings for engineering and technology (United Kingdom).

The rankings in Table 3 and 4 are broadly consistent with patent citations suggesting the technological value of a university. For example, the high positions of MIT and Cambridge are consistent with this, as are the relatively low positions of institutions that are not primarily known for science and technology, such as Harvard (14th), Columbia (18th), Princeton (21st), and in the UK the London School of Economics and Political Science (48th). There are some exceptions, however. Imperial College (3rd in the UK) is a UK university with a focus on applied science and industrial relevance, but with only 7 patent URL citations found by the crawler. This is partly due to its main domain name changing from ic.ac.uk (30 patent citations) to imperial.ac.uk and so the searches using its current main domain name would miss most patent citations to its older domain name. Nevertheless, even adding both searches would greatly underestimate its technological contribution and so it is an anomaly in the results. The Times Higher Education World University Rankings for engineering and technology in Tables 3 and 4 are a partial indicator of the strength of a university in one large patent-relevant area, and these suggest that the patent URL citation rankings are not very effective for the UK.

The number of academic URL citations to the top 5 US universities in Table 3 allows the creation of hyperlink networks connecting patents and university websites (Figure 1). Since one patent may link several universities (and any other online sources) these maps might permit the identification of technological topic relations between these institutions, regardless the original patent's assignee.

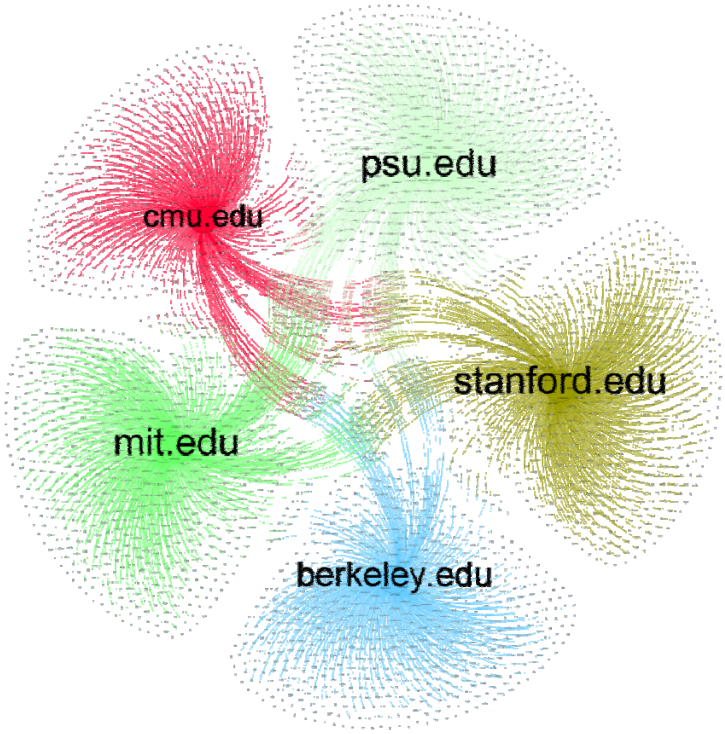

FIG. 1. Directed network of universities (MIT, Stanford, Berkeley, Carnegie Mellon University, Penn State University) and UPSTO patents.
Note: network created by Gephi (Fruchterman-Reingold algorithm)

RQ3: Types of impact reflected by patent URL citations

An investigation into the cited subdomains of the university websites can give insights into the reasons why university URLs occur in patents. The 28,900 patents with URL citations to at least one US University generated 132,567 URL citations, and for the UK the 2,667 patents generated 4,328 URL citations. The impact of special web services is evident in the most cited domains (Tables 5, 6). The ist.psu.edu domain dominates due to hosting the academic search engine CiteseerX (citeseerx.ist.psu.edu). Whilst this is successful site is a major achievement of the hosting university, it is cited for hosting the work of scholars from (predominantly) other universities and its URL citation count tends to dramatically overestimate the technological value of its hosting university. Similarly, the (now defunct) medical glossary at cancerweb.ncl.ac.uk, and the hosting of the *International Union of Biochemistry and Molecular Biology Recommendations on Biochemical & Organic Nomenclature, Symbols & Terminology etc.* hosted by Queen Mary (http://www.chem.qmul.ac.uk/iubmb/) have greatly inflated the counts for the hosting institutions through the work of others. Conversely, the liberal presence of computer science departments in the top positions suggests more home grown impact, but even in this case the dominance of computer science skews the rankings towards universities that specialise in applied computer science rather than any other forms of technology.

TABLE 5. Top 10 US university website domains cited in patents (crawler data).

Entity	University	US domain	URL citations
College of Information Sciences and Technology	Penn State	ist.psu.edu	13,860
School of Computer Science	Carnegie Melon	cs.cmu.edu	5,007
Media Lab.	MIT	media.mit.edu	3,779
Computer Science division	Univ. of California-Berkeley	cs.berkeley.edu	2,407

Dep. of Computer Sciences	Univ. of Wisconsin-Madison	cs.wisc.edu	2,125
Electrical Engineering & Computer Science Dep.	Univ. of California-Berkeley	eecs.berkeley.edu	2,005
The Robotics Institute	Carnegie Melon	ri.cmu.edu	1,912
Computer Science and artificial intelligence Lab.	MIT	csail.mit.edu	1,548
College of Computing	Georgia Institute of Technology	cc.gatech.edu	1,264
Dep. of Computer Science	Columbia	cs.columbia.edu	1,236

TABLE 6. Top 10 UK university website domains cited in patents (crawler data).

Entity	University	UK domain	URL citations
School of Biological & Chemical Sciences	Queen Mary, London	chem.qmul.ac.uk	304
The Computer Lab.	Cambridge	cl.cam.ac.uk	219
School of Informatics	Edinburgh	inf.ed.ac.uk	207
Cancer Web Project	Newcastle	cancerweb.ncl.ac.uk	173
Dep. of Engineering	Cambridge	eng.cam.ac.uk	153
School of Life Sciences	Sussex	lifesci.sussex.ac.uk	143
Dep. of Mathematical Sciences	Bath	maths.bath.ac.uk	122
Dep. of Electrical and Electronic Engineering	Surrey	ee.surrey.ac.uk	111
School of Informatics	Edinburgh	dai.ed.ac.uk	94
Dep. of Computer Science	Glasgow	dcs.gla.ac.uk	92

An alternative method to get insights into motivations for patent URL citations is to investigate highly cited web pages. This section examines the most cited pages in one major university for each country: Massachusetts Institute of Technology and the University of Cambridge (Tables 7 and 8). Since most are research articles, it seems that some academic publications are probably useful for technology transfer.

TABLE 7. MIT web pages with the most patent URL citations (crawler data).

Web page	Patents	Subject	Type	Available
people.csail.mit.edu/trevor/papers/1998-021/node6.html	430	Electrical engineering	Part of paper	Yes
people.csail.mit.edu/rywang/hand	358	Electrical engineering	Video demo	Yes
web.mit.edu/isn/research/team02/project02_04.html	348	Nanotechnology	Project information	No
people.csail.mit.edu/rywang/handtracking/s09-hand-tracking.pdf	341	Electrical engineering	Research article	Yes
web.mit.edu/m-i-t/articles/index-furniss.html	291	Physiology	Conference paper	No
pubs.media.mit.edu/pubs/papers/96_04-cmj.pdf	224	Media studies	Research article	No
pubs.media.mit.edu/pubs/papers/98-3-JNMR-Brain-Opera.pdf	211	Media studies	Research article	No
web.media.mit.edu	194	Media studies	Website	Yes
supertech.csail.mit.edu/papers/xaction.ps	129	Computer science	Research article	Yes
umech.mit.edu/6.021J/2004/lectures/lec06hi.pdf	124	Physiology	Research article	No

TABLE 8. University of Cambridge web pages with the most patent URL citations (crawler data).

Web page	Patents	Subject	Type	Available
vbase.mrc-cpe.cam.ac.uk	15	Biochemical engineering	Database	No
cl.cam.ac.uk/research/srg/netos/papers/2003-xensosp.pdf	14	Computer laboratory	Research article	Yes
cl.cam.ac.uk/research/srg/netos/papers/2004-oasis-ngio.pdf	12	Computer laboratory	Research article	Yes
cl.cam.ac.uk/research/srg/netos/xen/	10	Computer science	Virtual machine	Yes
mi.eng.cam.ac.uk/~bdrs2/papers/stenger_cvpr01.pdf	10	Engineering	Research article	Yes

mi.eng.cam.ac.uk/reports/svr-ftp/drummond_iccv2001.pdf	10	Engineering	Research article	Yes
cl.cam.ac.uk/research/srg/netos/papers/2005-migration-nsdi-pre.pdf	9	Computer laboratory	Research article	Yes
mi.eng.cam.ac.uk/research/vision/research.html	9	Engineering/Computer science	Research material	No
cl.cam.ac.uk/~mgk25/unicode.html	8	Computer laboratory	Information resource	Yes
cl.cam.ac.uk/~sk507/pub/04-cmg-JBoss.pdf	8	Computer science	Conference paper	No

RQ4: General patent URL citations

The websites and web pages most cited by patents form an alternative source of information about common purposes for patent URL citations (Table 9). Although research information sites are prominent in the results (ieeexplore.ieee.org, citeseerx.ist.psu.edu, research.microsoft.com), as are computing-related information sites (w3.org, schema.org, tools.ietf.org), Wikipedia and the Internet Archive Wayback Machine for old (or more permanent) URLs (web.archive.org). Wikipedia, the Wayback Machine and YouTube sites are general enough to give little insight into why they are used. YouTube seems to be frequently cited for sales or review videos demonstrating the features of competing products as a convenient way of describing them to avoid the need for textual descriptions of designs (e.g., https://google.com/patents/USD700164) or to demonstrate the features of a product (e.g., http://www.google.co.uk/patents/USD645468).

TABLE 9. Most common domains from URL citations in patents citing the selected US and UK universities (crawler data).

Domain (US)	Patents	Domain (UK)	Patents
youtube.com	36,256	youtube.com	3,483
schema.org	28,900	schema.org	2,667
web.archive.org	24,363	en.wikipedia.org	1,241
ieeexplore.ieee.org	20,750	citeseerx.ist.psu.edu	1,007
en.wikipedia.org	17,851	ieeexplore.ieee.org	892
citeseer.ist.psu.edu	13,770	web.archive.org	805
research.microsoft.com	10,330	research.microsoft.com	777
w3.org	6,205	oblong.com	748
msdn.microsoft.com	4,937	fingerworks.com	721
oblong.com	4,836	tools.ietf.org	609

The most cited Wikipedia pages seem to be for quite technical definitions or descriptions, often with computing-related themes (Table 10). Thus, the presence of Wikipedia tends to confirm the computing bias of URL citations.

TABLE 10. Most cited Wikipedia pages from the US set of patents (crawler data).

Wikipedia Term	Frequency
en.wikipedia.org/wiki/Wi-Fi	419
en.wikipedia.org/wiki/Eye	401
en.wikipedia.org/wiki/Vocative	373
en.wikipedia.org/wiki/Image	299
en.wikipedia.org/wiki/Automatic	248
en.wikipedia.org/wiki/Optical	244
en.wikipedia.org/wiki/Exception-Handling	242
en.wikipedia.org/wiki/Waypoint	214
en.wikipedia.org/wiki/Hough-transform	200
en.wikipedia.org/wiki/Wired	176
en.wikipedia.org/wiki/Cosmic-ray	156
en.wikipedia.org/wiki/Speculative-execution	156

en.wikipedia.org/wiki/BGM-109	149
en.wikipedia.org/wiki/Bayesian_inference	148
en.wikipedia.org/wiki/Bayes-theorem	148
en.wikipedia.org/wiki/Bayesianism	140
en.wikipedia.org/wiki/Bayesian	133
en.wikipedia.org/wiki/Bayes	130
en.wikipedia.org/wiki/software	128
en.wikipedia.org/wiki/Cosmic	122

Limitations

A limitation of the method used is that the HTML patents are often generated using OCR scanning, which generated a number of errors in the URLs. For example, there were at least 49 incorrect variants of web.archive.org, such as web.arcbive, web.archive.crg, and web.archve.org. Thus, without manual cleaning the patent URL citation counts are likely to be underestimates, whichever method is used.

Another limitation is that the results include an unknown proportion of self-citations. Including these may inflate the counts for universities with many patenting inventors, assuming that inventors disproportionately cite their own contents. Nevertheless, such citations are still useful evidence of technological value produced by the host university.

An important issue is that only two countries were analysed in the current paper and the number of patent URL citations to universities in most other countries is likely to be substantially lower. Thus the results should be viewed as the best case scenario rather than as a typical case. This problem would probably not be solved by analysing patents from other countries since the US patent system seems to use citations more extensively than others.

A final limitation is that the patent URL citations have not been systematically analysed in the context of the citing patent and the reasons for their creation have been broadly inferred from the cited URLs and cited web domains. In many cases it is difficult to be sure why a non-patent citation has been added to a patent because it may only form part of a bibliography of related documents rather than being cited within the text and so there is no straightforward way to identify explicit reasons for patent URL citations.

Discussion and conclusions

The occasionally inflated results from the SERP100_1 method confirms the inconsistency of the Google Patents search results when there are many hits. This simple method is probably only useful for the purpose of ranking major US or top UK universities. Although the most accurate method to get patent URL citation counts is probably to crawl the Google Patents website and then extract the URLs from the downloaded pages, both SERP100_n and SERP10_n seem to be reasonably accurate and are much simpler practical alternatives.

Nevertheless, it seems unlikely that there are enough patent URL citations to give useful data for most of the world's universities. Moreover, the results are biased towards universities with strong computer science and can be greatly influenced by university domains that host widely used information resources, so the patent URL citations do not always reflect knowledge developed at the cited university. Overall, then, patent URL citations are not a strong new indicator, except for major US and the top few UK universities, and should be filtered for irrelevant influential web domains, if used. Other technology transfer indicators, such as the number of patents granted to a university or

the number of traditional citations in patents are probably more useful in practice, however.

In the wider context of academic-related link analyses, even though patent URL citations seem to avoid spam and irrelevant content, their value is undermined by other problems and the low numbers found overall. Whilst some social web article-level metrics have been adopted by publishers and are marketed by companies such as altmetric.com, link-based metrics are still primarily useful only for the Ranking Web of World Universities.

However, although the number of web citations in patents targeting university websites is limited, the use of general URL citations is significant. A total of 700,815 hyperlinks have been captured from the US sample (28,900 UPSTO patents), reflecting the widespread use of patent web citations to general-purpose sources such as Wikipedia, YouTube and the Internet Archive.

This paper has introduced the idea of URL citation analysis for patents. The different methods tested provide a deeper understanding about the different methods and indicators that are suitable for analysing full-text digitised patents. This opens the way to future research on the connections between inventors/inventions and cited online resources, and represents a preliminary and exploratory first step towards understanding the citing of web content from patents.

References

Aguillo, I. F., Ortega, J. L., & Fernández, M. (2008). Webometric ranking of world universities: Introduction, methodology, and future developments. *Higher education in Europe*, 33(2-3), 233-244.

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

Banks, D. (December 13, 2006). Now you can search for U.S. patents. Google Official Blog. Retrieved from http://googleblog.blogspot.com/2006/12/now-you-can-search-for-us-patents.html

Bar-Ilan, J. (2004a). A microscopic link analysis of academic institutions within a country – The case of Israel, *Scientometrics*, 59(3), 391–403.

Bar-Ilan, J. (2004b). The use of Web search engines in information science research. Annual Review of Information Science and Technology, 38(1), 231-288.

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management*, 41(3), 973–986.

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107–117.

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

Endres, J. (2007). Take Google Patent Search for a spin. INFORM: International News on Fats, Oils and Related Materials, 18(6), 400.

Harries, G., Wilkinson, D., Price, E., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science*, 30(5), 436–447.

Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, 5(3), 446-457.

Ingwersen, P. (1998). The calculation of web impact factors. Journal of Documentation, 54(2), 236-243.

Kim, H. J. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science*, 51(10), 887–899.

Kousha, K, & Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. *Scientometrics*, 68(3), 501-517.

Kousha, K., & Thelwall, M. (in press). Patent citation analysis with Google. *Journal of the Association for Information Science and Technology*.
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.
Marley, M. (2014). Full-text patent searching on free websites tools, tips and tricks. *Business Information Review*, 31(4), 226-236.
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, 17–27.
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.
Moskovkin, V. M., Shigorina, N. A., & Popov, D. (2012). The possibility of using the Google Patents search tool in patentometric analysis (based on the example of the world's largest innovative companies). *Scientific and Technical Information Processing*, 39(2), 107-112.
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 US 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.
Orduna-Malea, E. (2013). Aggregation of the web performance of internal university units as a method of quantitative analysis of a university system: The case of Spain. *Journal of the American Society for Information Science and Technology*, 64(10), 2100-2114.
Orduna-Malea, E., & Ontalba-Ruipérez, J. A. (2013). Selective linking from social platforms to university websites: a case study of the Spanish academic system. *Scientometrics*, 95(2), 593-614.
Orduna-Malea, E., Torres-Salinas, D., & Delgado López-Cózar, E. (2015). Hyperlinks embedded in twitter as a proxy for total external in-links to international university websites. *Journal of the Association for Information Science and Technology*, 66(7), 1447-1462.
Ortega, José L., Orduna-Malea, E., & Aguillo, Isidro F. (2014). Are web mentions accurate substitutes for inlinks for Spanish universities?. *Online information review* 38(1), 59-77.
Park, H. W. (2002). Examining the determinants of who is hyperlinked to whom: A survey of Webmasters in Korea. *First Monday*, 7(11). Retrieved from http://www.firstmonday.org/issues/issue7_11/park/index.html
Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/2874
Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. Cybermetrics, 2/3, http://www.cybermetrics.info/articles/v2i1p2.pdf.
Seeber, M., Lepori, B., Lomi, A., Aguillo, I., & Barberio, V. (2012). Factors affecting web links between European higher education institutions. *Journal of informetrics*, 6(3), 435–447.
Smith, A. G. (1999). A tale of two web spaces: comparing sites using web impact factors. *Journal of documentation*, 55(5), pp.577–592.
Smith, A. G., & Thelwall, M. (2002). Web impact factors for Australasian universities. *Scientometrics*, 54(3), 363-380.
Sud, P., & Thelwall, M. (2014). Linked title mentions: A new automated link search candidate. *Scientometrics*, 101(3), 1831-1849.

Thelwall, M. (2001). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52(13), 1157–1168.

Thelwall, M. (2002). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3). Retrieved from http://www.informationr.net/ir/8-3/paper151.html

Thelwall, M. (2004). *Link analysis: an information science approach*. Amsterdam: Elsevier.

Thelwall, M. (2008). Quantitative comparisons of search engine results, Journal of the American Society for Information Science and Technology, 59(11), 1702-1710.

Thelwall, M., & Kousha, K. (2015). Web indicators for research evaluation. Part 1: Citations and links to academic articles from the Web. *El profesional de la información*, 24(5), 587-606.

Thelwall, M., & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organisations. *Journal of the American Society for Information Science and Technology*, 62(8), 1488-1497.

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.

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.

Vaughan, L. (2015). Uncovering information from social media hyperlinks: An investigation of twitter. *Journal of the Association for Information Science and Technology*, 67 (5), 1105-1120.

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.

Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science*, 29(1), 59–66.