Can the Web Give Useful Information about Commercial Uses of Scientific Research?

Mike Thelwall[1]
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478

Invocations of pure and applied science journals in the Web were analysed, focussing on commercial sites, in order to assess whether the Web can yield useful information about university-industry knowledge transfer. On a macro level, evidence was found that applied research was more highly invoked on the non-academic Web than pure research, but only in one of the two fields studied. On a micro level, instances of clear evidence of the transfer of academic knowledge to a commercial setting were sparse. Science research on the Web seems to be invoked mainly for marketing purposes, although high technology companies can invoke published academic research as an organic part of a strategy to prove product effectiveness. We conjecture that invoking academic research in business Web pages is rarely of clear commercial benefit to a company and that, except in unusual circumstances, benefits from research will be kept hidden to avoid giving intelligence to competitors.
Keywords: Web mining, web searching, webometrics, business intelligence, knowledge transfer

Introduction

The nature of science research has evolved over time, from early isolated amateur practitioners to organised teams (Gross, Harmon & Reidy, 2002), interspersed with large-scale 'big science' enterprises (Price, 1963). One recent trend in academic research is for an increasing focus on commercial problem-solving in large multi-disciplinary teams (Gibbons et al., 1994). This is an issue for government, which plays important roles in the collaboration between universities and industry, including as a paymaster of academia, and needs to ensure that results of some academic research transfers efficiently to commercial settings (e.g. Etzkowitz & Leydesdorff, 1997; Moncada, Rojo, Bellido et al., 2003). Scientific publications have traditionally been a convenient means for government and others to evaluate the quality of scholars' work (e.g. van Raan, 2000), if part of a wider framework (Tijssen, 2003), but the new problem solving teams may deliver other outcomes such as novel products or production processes. If they publish their results, natural outlets would be applied and professional journals that may not be held in high esteem by the research community. Those involved with the study and assessment of science have sought novel information sources with which to monitor the new style science, with a natural contender being the Web since it seems to be standard practice for companies to maintain web sites (at least in the richer nations). Liu and Arnett (2000) identify information provision as a key factor in the success of e-commerce web sites, and so it seems possible that commercial web sites will often contain information about academic partnerships or uses of research. Exploratory research is therefore necessary to assess whether the Web can yield useful information concerning university-industry knowledge transfer.

A related second issue that the Web may also be able to address is whether applied research journals have an impact that is consistently underestimated by the journal Impact Factors of the Institute for Scientific Information (ISI), used for many different levels of research evaluation (Moed, 2002). If applied journals had a relatively higher citation count on the non-academic Web (however operationalized) than theoretical journals from the same field then this would provide confirmatory quantitative evidence.

[1] To appear in Online Information Review, 2004, Vol. 28 part 2, pp. 120-130.

Assessing University-Industry Knowledge Transfer

A wide variety of metrics have been developed to investigate or monitor various aspects of academic or business operations (Rubenstein & Geisler, 1991; Geisler, 2000; Schmoch, 2003) and to assess innovation. Historically, peer review and citation analysis have dominated academic research quality assessments, but recently the emphasis has shifted towards seeking evidence of applicable knowledge production, such as the receipt of grants from external sources. Patents, however, are widely used for the assessment of technological progress and seem to be particularly promising for studies of university-industry relationships (Meyer, 2000; Meyer, 2003; Oppenheim, 2000; van Looy, Zimmermann, Veugelers et al., 2003). Patents sometimes have reference lists citing other patents and academic journal articles, which allows the identification of "those fields of technology that are highly science interactive" (Verbeek, Debackere, Luwel et al., 2002), but the extent to which commercial web pages contain such information is unknown. Simple methods of assessing the commercial potential of large research groups include patent grant counts. Oppenheim (2000) draws attention to the need for different kinds of validation studies to assess the value of patent citation analysis and allow patent statistics to be assessed with confidence. The same is true by extension for any web-based approach with similar motivations.

Although much web research has focused on hyperlinks as analogous to references in journal articles (Ingwersen, 1998), invocations (mentions in the text of a page) are a more promising information source since they should be more numerous. For example, a journal may be mentioned in a web page without a link to its web site, but such a link could be expected to be accompanied by the name of the journal in or near the link description text. Some previous studies have focused on web invocations in a scholarly context. Cronin et al. (1998) analysed web invocations of scholars' names, finding a wide range of contexts although none that were explicitly commercial, presumably because the Web at the time (1997 or early 1998) was dominated by academic content. Landes and Posner (2000) use web invocations of scholars in conjunction with two other sources: media mentions and citation counts, in order to gain wide evidence of "public-intellectual status" (Posner, 2001). Vaughan and Shaw (2003) have investigated invocations of Library and Information Science journal articles in the general web, finding that counts of web invocations of articles in most (57%) of the journals correlated with citations counts. For whole journals, the online invocation counts correlated significantly with the Institute for Scientific Information's Journal Impact Factors. This suggests that web invocations of journals may yield information about research impact, perhaps including university-industry knowledge transfer. Leydesdorff and Curran (2000) have conducted a comparative analysis of online connections between university, industry and government sites but did not investigate the phenomenon of technology transfer.

What is known about why journals are invoked in web pages? There is a long history of investigating the related topic of citer motivations in journal articles and recently similar questions have been asked about electronic phenomena such as e-journal article citations (Smith, 1999; Kim, 2000; Harter & Ford, 2000) and general web links (Thelwall, 2003; Wilkinson *et al*., 2003). These have typically found a very wide set of motivations, reflecting both the extra capabilities of the electronic environment and the wider types of context in which information is presented on the Web. Invocation motivations on the Web are likely to include new ones because of the types of publication on the Web that are publicly available for the first time, including course reading lists and commercial information provision web sites (Sloan, 2001; Thelwall, 2002a).

Research Design

In order to decide whether the Web could be used as an information source about university-industry knowledge transfer, academic research was operationalized as the contents of academic journals and information about university-industry knowledge transfer was sought through the study of commercial web pages that invoke academic journals. This is clearly an oversimplification and also does not answer the need to gain information about academic

research that does not result in journal articles. The goal was not to obtain a new all-encompassing data source, however, only one that contains *some* useful information about commercial exploitation of academic research. The two specific research aims were (a) on a macro level to discover whether the web-based invocation impact (counts of text mentions) of applied journals was relatively higher than the Institute for Scientific Information's (ISI) corresponding journal Impact Factors and (b) on a micro level to discover what types of evidence could be found in commercial web sites about university-industry knowledge transfer. Scientific research is the scope of the study, because outside of the sciences research is often transmitted primarily in books rather than journal articles (Hyland, 2000; e.g. Vann & Bowker, 2001), and it would be much more difficult to identify lists of relevant book titles than relevant journal titles.

The overall research design was to select two scientific disciplines and to analyse two collections of journals for each one; a collection with a more theoretical orientation and one with a more applied orientation. The commercial search engine Google was used to count and identify as many web pages as possible that invoked any of the journals, and a human classifier coded a random selection of the pages. Google was chosen for its large web coverage in addition to its indexing of some non-HTML formats, and Portable Document Format (PDF) and PostScript (PS) in particular. The counts produced would address aim (a) and a breakdown of the results of the classification would then reveal the range of common invocation contexts and whether significant differences were present or not (aim (b)). A similar exercise was conducted by a second classifier for triangulation purposes. The combination of qualitative and quantitative approaches makes this a mixed model methodology (Tashakkori & Teddlie, 1998).

Note that the use of a search engine for the initial set of pages invoking journal articles is unavoidable for projects aiming to cover the 'whole Web' (Thelwall, Vaughan & Björneborn, 2005), but is problematic because search engine coverage of the Web is demonstrably partial (Lawrence & Giles, 1999). The lack of complete web coverage is a conceptual issue that can be avoided by acknowledging that the scope of the study is only Google indexed pages. These will tend to be the more important pages on the Web (Brin & Page, 1998) and the ones that are most likely to be found by users because of the popularity of Google, and so this is a practical and reasonable scope. Other undesirable and unavoidable factors associated with search engine use for research, include a lack of control over, and knowledge of, the algorithms that collect and report the data (Bar-Ilan, 2001; Björneborn & Ingwersen, 2001), but their successful exploitation in the past justifies their continued use, (e.g. Thelwall, 2002b) provided that results are interpreted cautiously.

Methods

Field and Provisional Journal Selection

The first task was to select comparable pure and applied science fields so that the online impact of their journals could be compared. The categories used by the Institute for Scientific Information (ISI) were used as the basis because these are a time-tested and influential third-party source of information. Chemistry and Physics were selected from this list of fields because there were several different subcategories for each, including apparently pure categories and apparently applied categories in both cases. Chemistry, Applied and Physics, Applied were provisionally chosen as the applied categories and Physics, Mathematical and Chemistry, Inorganic as the pure. We discussed the selection with subject experts and although the Physics choices were unproblematic, the chemists felt that all the non-applied Chemistry categories contained a mix of pure and applied research. We eventually chose Chemistry, Physical as the most pure category whilst recognising it to be mixed pure and applied rather than pure.

The Web of Science (ISI, 2003) was used to obtain a listing of journals in all four categories. Some journals appeared in two different categories, and these were removed from the pure/mixed category. These were assumed to be applied in nature because of being in an

explicitly applied category, and so their removal would make the pure/mixed categories purer. Each journal name was converted to its full title from the ISI abbreviation. Journals with "applied" in the title but in a pure or mixed list were removed.

Random URL Selection

A count and list of web pages that mentioned each journal name was obtained from Google by entering the journal title as a phrase search. The first results page for each journal was then examined for false matches. Journals that recorded at least one false match on the first page were dropped from the list. The typical cause was a common name or the existence of another journal with a longer name, containing the first name as a substring. The purpose of this filtering was to exclude journals for which the Google results were likely to have a significant degree of error. At this stage journals with a majority of citations in non-English web pages were also excluded. The reason for this is that there were relatively few of them and their inclusion would have potentially skewed the results if there were substantially different national-linguistic patterns of journal invocation. This seems likely because of the existence of national bodies of scholarly literature, sometimes language-specific (van Leeuwen, Moed, Tijssen, et al., 2001).

For the remaining 123 titles, Google was used to fetch the first 1,000 matches. The default search option in Google is to hide "similar matches" which means showing only one or two pages per site. It is possible to turn off this option in searches, which would have given the same number of hits but from a smaller range of sites. The default option was kept in order to limit the effect of duplicate or near-duplicate reasons for including the journal name, assuming that citations from the same site are more likely to originate from a common reason than those from different sites. Conceptually similar procedures, such as the Alternative Document Models, have been previously used in web research (Bharat, Chang, Henzinger & Ruhl, 2001; Björneborn, 2001; Thelwall, 2002a; Thelwall & Wilkinson, 2003a). This is clearly a heuristic and not a perfect solution, representing a practical use of the tools available. This approach will be termed the Google Virtual Document Model (VDM). The set of up to 10 results pages for each journal each citing up to 100 pages containing journal name was fetched from Google (manually) and saved to disk. The URLs were stripped out of each page by a program and compiled into a single list of 74,446. Note that the Google API (Google, 2003) could have been used to automate the initial collection of the results pages, which should have given the same outcome.

The next step was to remove library style sites and other long inclusive lists of journals from the URLs. The reason for this was that many URLs retrieved were simply lists of journals from library sites and appeared in many of the journal searches, having an undue influence on the results. URLs failing the following heuristically designed 'replication' test were therefore removed, leaving 42,540.

- Any with 5 or more URLs having identical paths up to the start of a *query* i.e. up to the first question mark in the URL
- Any with 5 or more URLs having identical paths up to the final *directory,* when a query was not present
- Any with 11 or more URLs having identical *domains*

The reduced list of URLs formed the raw data for the research from which the random sample was selected for classification. Before selecting the sample, however, two further steps were taken. First, the URLs were split into university and non-university groups. URLs were identified (by a program) as being from a university if their domain names either contained a recognised academic designation: ".edu" or ".ac." or matched the domain name of one of the 4,360 universities in the online list at `http://geowww.uibk.ac.at/univ/world.html`, excluding universities not having their own domain name. Note that government funded research institutes that are not universities were effectively classified as non-academic. This includes organisations such as the Max Planck Institute in Germany and the Council for Scientific Investigations (CSIC) in Spain. The precise decision about which to include is problematic because the direction of

government funding for research varies by country, with some mainly funding research inside universities, whereas others support external non-commercial bodies, or even encourage private initiatives (New Scientist, 2003).

The final list of 4,000 URLs to be categorised was compiled from a random selection of 500 university and 500 non-university URLs from each of the four journal categories. The even split was designed to stop academic sites from swamping the sample, so that the categories could be compared, and the most important category, commercial sites (which could not be automatically identified) would be present in reasonable numbers. Each URL was downloaded to disk using the WinHTTtrack software and two CD-ROM copies made, one for each classifier. The URLs were numbered in a random order so that the validity of the results would not be undermined by the classifiers not completing the whole set. Note that Google applies hidden techniques based around PageRank (Brin & Page, 1998) to choose which pages it places in the top 1000 matches lists that were accessed by the method used. This potential source of bias in the data set was not a big problem in practice, since with the VDM, most journals matched less than 1,000 URLs anyway.

The proportion of URLs invoking each journal but failing the duplication test and not being flagged as education-related was also recorded for use in estimating the total number of non-replicated non-academic links as an indicator of online non-academic journal impact (aim (a), used for Table 1).

The Classification Scheme

The objective of the classification scheme was to find out why journal names were invoked in web pages, particularly those outside of educational contexts (see Appendix). The scheme was designed to classify the context of each invocation in terms of what was invoked, what type of organisation owned the invoking page and what type of page it was. A pilot scheme was drawn up based upon a previous exercise (Wilkinson et al., 2003) and tested by the author (using URLs from outside of the 4,000 selected) and refined, principally by adding extra categories for newly observed phenomena. The scheme included categories and reasons why the category could be selected, modelled upon content analysis (Krippendorff, 1980). A training exercise was then conducted with two paid classifiers on another test set and the categories and reasons for choosing categories again refined. The two classifiers were then given the full data set to classify independently over a period of three months. The classification differs from content analysis in that it was accepted from the start that the data would be difficult to classify and that it would be impossible to get a high degree of inter-classifier agreement. The second classifier therefore serves the purpose of triangulation of the *overall* results rather than confirmation of the individual codes.

Results

Impact Factors and Invocations

Table 1 gives the median values of the ratios of invocation counts to journal Impact Factors. Medians are appropriate for the data because it seems likely that it will be fundamentally of a skewed nature, as typically are hyperlink counts (Barabási & Albert, 1999), and word frequencies in text data (Zipf, 1932). The first column of numbers reports the raw invocation counts returned by Google. The second column gives the (median across journals) proportion of URLs in the set returned by Google that were automatically identified as being either replicated (in the sense described above) or from a university site. The proportions were calculated separately for each journal and used as an estimator for the proportion of URLs in the complete set (i.e. including URLs counted by Google but not returned in its results) that were of non-educational, non-replicated origin. The final column reports the median invocations having adjusted for the unwanted page sources. This represents a very approximate estimate of the non-academic invocation impact of the journals. Note that the proportions reported in Table 1 will be are almost certainly overestimates because the pages checked all come from the Google VDM, which tends to hide the more replicated pages. No

statistical significance tests were conducted because the results would be misleading given the assumptions and approximations used.

Table 1. Invocation counts compared to impact factors.

Area	Median Invocations JIF	Median proportion of non-education non replicated pages in the returned set	Median estimated non-replicated non-educational invocations JIF
Chemistry, Applied (chemapp)	1518	0.28	360
Chemistry, Physical (chemphys)	1441	0.13	187
Physics, Applied (physapp)	2251	0.18	460
Physics, Mathematical (physmath)	4029	0.15	528

Classification of Invocations

The classification results were separated into 12 categories, one for each of the journal types and each of the identified types of owning organisation (university, government, industry). The results for the owning organisations were very similar to each other and so to simplify the process of reporting the findings, only the results for the key group is graphed, that of commercial pages, and the other results will be alluded to only when they differ substantially. The main classifier categorised 2,450 pages, with 335 of them identified as definitely commercial and so the primary results below relate to these 335 pages.

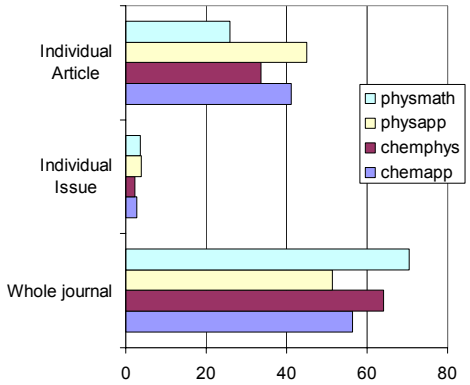

Figure 1: A breakdown of types of commercial invocation targets (%).

What is invoked (Figure 1): Whole journals are the most commonly cited overall with slightly less individual articles and almost no individual journal issues. The pattern for university pages is similar, but with the Physics, Mathematical pages being evenly split between individual articles and whole journals. Government pages cite more individual articles than whole journals in each of the four categories, reversing the trend above for commercial companies.

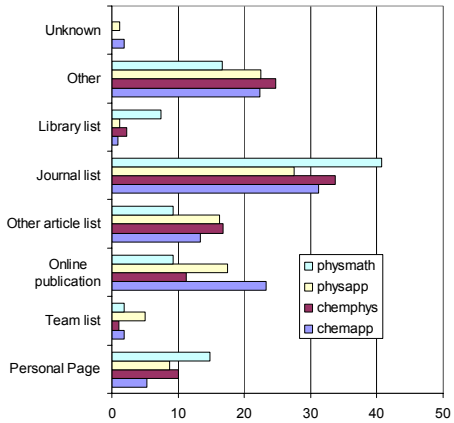

Figure 2: Commercial invocation sources (%).

Invocation sources (Figure 2): Most citations came from simple lists of journals, mainly outside of library sites, despite the most highly replicated urls having been previously filtered out. A few came from list of publications by an individual or team. Online publications such as journal or conference articles accounted for a significant number too, with these being more in evidence in applied than pure areas. Education page types were similar in spread except for more personal pages (28% overall). Less than 2% of course pages invoked journal titles in the educational pages, in contrast to the 12% found for Library and Information Science by Vaughan and Shaw (2003), which may reflect a lower use of research in teaching in the hard sciences. The spread of invocation sources in government pages also showed a similar pattern with a higher percentage of personal pages (21% overall).

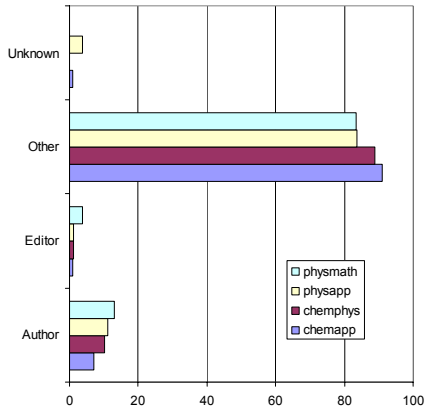

Figure 3: Who invoked the journal from a commercial site (%).

Invoking page owner (Figure 3): Author invocation of their own publication was rare in commercial web sites, but more common in Education and Government Web pages, accounting for an average of 27% and 24% of invocations respectively.

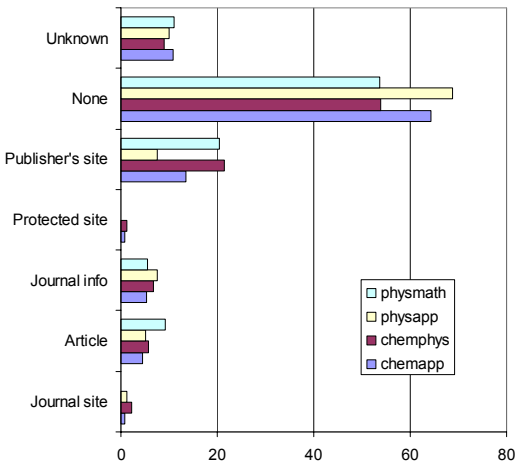

Figure 4: Commercial site links associated with journal invocations.

Links (Figure 4): In most cases there was not a link associated with the invocation, but when there was, it commonly targeted to a publisher's site for the journal. A few links also targeted the article, general information about the journal, or a password protected site.

Triangulation

The data was processed for the second classifier as above and the same descriptive statistics calculated (not shown). The results of the two classifiers were compared and found to be very similar overall, with the differences between the results of each category being typically less than 5%. The main exception was that the second classifier hardly used the category 'Other article list' in Figure 3, using 'Journal list' as a replacement for it (see Appendix for the category descriptions). Also, in the data for Figure 4, the second classifier tended to identify less links. Overall however, the differences found do not fundamentally undermine the validity of the results as reported above, with the exception of the 'Journal list' and 'Other article list' categories. The difference between these two should be interpreted with caution. With this exception the triangulation establishes that the figures presented are not pathologically unreasonable interpretations of the categories.

Classification of Uses of Academic Research in Commercial Settings

The mentions of journals in commercial web pages were investigated to see whether there was clear evidence of academic research being used in a commercial setting. Of the 335 pages invoking a journal article identified as commercial by the first classifier, 37 (i.e. 11%) were identified as examples of academic research cited in a context where the research was being helpful to a commercial company. The identification was made primarily by the two classifiers, with the list compiled by both of them being merged and verified by the author. These were classified by the author using an inductive process - grouping apparently similar invocation contexts and choosing a label for the group - and the results reported in Table 2.

Table 2. A classification of academic research invoked in contexts helpful to business.

Description	No.	Example
Trade magazine or trade organisation site	10	http://www.nzwine.com/assets/Health_Benefits_Review.pdf Report to New Zealand winegrowers on the benefits to health of moderate wine consumption
List of own publications	10	http://www.hll.com/HLL/careers/articles.html The publications of Hindustan Lever Laboratories
Prove product properties	8	http://www.polyone.com/ind/doc/vinyl_floor.asp A polymer services company promoting the qualities of vinyl flooring
Papers published using the company's product	5	http://www.metacomptech.com/cfd++/01-3021.pdf An article that employed the software made by the company hosting the paper
Research used in making a company's product	2	http://www.mtiresearch.com/nld/refspubs.html "We have found the following list of references valuable in understanding the general theories of nonlinear dynamics…" http://www.lumera.com/html_only/whitepaper.html "…published papers that discuss the foundation of Lumera's technology."
List of academic research to boost company credibility	2	http://www.dataphysics.de/deutsch/service_lit.htm "our selection of current standard literature" about surface and colloid chemistry

Discussion

The goal of this research was to discover whether the Web could be used as a source of information about university-industry knowledge transfer. The first results (Table 1), comparing ISI impact factors to a web invocation equivalent found that applied journals were not invoked online relatively more frequently than pure and so on a large scale this is not the case. Despite commercial content apparently dominating the Web (Lawrence & Giles, 1999) this appears to be partly the result of the majority of the invocations occurring either in university sites or in large lists of journals, which increases the total invocation count of the applied journals more than that of the pure in both disciplines. After applying an heuristic to remove replicated invocations and using a list of university names to remove the university sites, the remaining estimated invocation counts should reflect more the non-academic impact of the journals. The revised figures (the final column in Table 1) did show more non-academic invocation impact, relative to traditional journal impact, for Applied Chemistry relative to Pure, but the results for the two Physics fields were very similar. The results are therefore partly ambiguous. First, the raw Google journal invocation counts do not seem to be of any value in identifying wider impact for applied scientific research. Second, if steps are taken to automatically filter the data, as above, then the results may be useful in some fields of science but not others. Confidence in the use of such statistics, even for Chemistry with its clear numerical differences, is undermined by the approximations and heuristics needed to produce it, and so it is doubtful whether this can be genuinely convincing. If Physics proves to be an anomalous case, however, with other areas of science conforming to the Chemistry model, then the overall case would be much more believable.

Figures 1 to 4 paint a picture of the context of invocations for commercial web sites. Recall, however, that the statistics exclude the almost half of pages that originate in large library or list sites. It is very common for entire journals to be invoked, rather than articles, and often in pages containing lists of relevant journals. Some practices that could perhaps be described as of academic origin are also present, for example the listing of the publications of an employee or research team. This could be self-publicity as part of a career advancement strategy (Hyland, 2003) or in a research-based company it could have the purpose of marketing the company's research skills. Possibly the invocation count-impact factor comparisons would have given more significant results if the invocation of whole journals had been excluded, but this would have been a very time-consuming task.

Perhaps the clearest outcome was the sparseness of genuine applications of academic research, as reflected in the 11% of invocations classified as directly helpful to the company, and the breakdown of categories in Table 2. It is rare for companies to publish the academic origins of their products or services, presumably to avoid informing their competitors (e.g. Liu, Ma & Yu, 2001). The exceptions seem to form two main groups. The first are high technology companies selling complex products that can use published academic research as a quality guarantor. This is a particularly interesting concept because it involves transferring knowledge from the commercial to the academic domain, the reverse direction to that being investigated here. The second exceptions are where academic research is used as a marketing tool to support contentions about the properties of a product. This has similar purpose to the first group but is an example of academic domain knowledge helping business, although not in product creation. In summary, it seems that academic research can be usefully cited in some types of commercial web site, but only to directly support sales strategies. Variation in Web site use is not surprising: even large similar companies have greatly differing online strategies (Perry & Bodkin, 2002). The issue of trust on the Web is critical, particularly for small businesses, and hyperlinks between organisations can help users to trust the sites (Stewart, 2003) (as earlier predicted by Davenport and Cronin (2000)). In this context, allusions to academic research can be seen as another trust-generating strategy, although almost certainly a less frequently employed one.

Disciplinary and pure/applied differences are difficult to comment upon because although there were differences in most of the graphs, some large, they would need to be more systematic to gave any confidence that they were not peculiar to the particular journal sets chosen and not affected by factors endemic to the Web that undermine the independence of the data set, such as the copying of web pages within and across sites (Chakrabarti, 2003). Nevertheless, it appears that there are real differences between similar fields in the extent to which their journals are invoked online. These may have many different causes, including established cultural practices, differences in the size of the field or spread of researchers, or differences in the use made of the field by other fields. A theoretical orientation that maybe significant is for a greater proportion of applied journals to be invoked in online articles in commercial sites, with pure journals being invoked relatively more frequently in journal lists and personal pages. Also, a greater proportion of applied invocations were journal articles rather than journal titles, perhaps indicating more direct use.

The different sources of links, both education and government, did show differences from commercial pages. More individuals' pages were found outside the commercial sites, perhaps reflecting a more marketing-oriented focus to the latter. Presumably commercial sites are typically more strictly controlled, a cultural difference in creation strategies and perhaps also concepts of ownership. This suggests that different kinds of academic-related information could be mined from each of them.

Journal invocations were analysed in our study as an operationalization of academic knowledge. The evidence now points to invocations being sometimes a source of evidence about the value of applied research but rarely strong evidence in the sense of clearly demonstrating technology transfer. Could different operationalizations have given better results? If journal *article* invocations had been used, then this would have produced improved results for the impact-invocation comparison (see Figure 1), perhaps enough to give applied physics journals a relatively higher online impact than pure physics. The problem with article identification is the extra resources needed to identify and track the individual articles from journals. This could be automated by combining the Google API with electronic sources of journal article names, however, and this direction is worthy of future research.

Talking a step further back, could there be significant non-journal web based evidence for university-industry knowledge transfer? For example, perhaps companies credit universities by name for effective collaboration, but do not mention any published research. Extrapolating from the cases examined here, it seems likely that such intelligence would not normally be published online for fear of helping competitors, but could occasionally be used as part of image promotion, product quality validation or marketing activities. This would probably be relatively infrequent and perhaps also specific to particular market sectors.

Certainly it now seems unlikely that companies would as a general rule see the need to explicitly acknowledge successful university-industry transfer. It is still probably the case, then, that the web will provide islands of evidence about university-industry knowledge transfer, but nothing like widespread coverage.

Conclusions

The results of the classification exercise suggest that the Web is not going to be a source of high quality macro level information about university-industry knowledge transfer, although there is some potential for gaining numerical evidence of the interest in academic journals by non-educational organisations. The results indicated that this was a possibility through a substantially higher invocation count for applied chemistry journals than for the more pure chemistry journals, although this was not evident for Physics. Existing numerical techniques for evaluating university-industry knowledge transfer, such as patent analysis and co-authorship analysis, are unlikely to be threatened by web-based approaches but may be complimented by the information found on the web if automated tools are developed to make the basic tasks of data collection and filtering not onerous.

It seems that it is rarely beneficial for a company to signal its use of academic research on its web site, and may often actually be harmful in terms of providing useful intelligence to competitors. Despite this, in a small number of cases academic research does seem to have a place in commercial web sites. The most common uses seem to be through demonstrating the credentials of the company or a specific product, a reinforcement of the importance of trust on the Web.

Acknowledgements

This research was funded by a grant from Emerald, publisher of academic and professional literature.

References

Barabási, A.L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.

Bharat, K., Chang, B., Henzinger, M., & Ruhl, M. (2001). Who links to whom: Mining linkage between web sites. In: IEEE International Conference on Data Mining (ICDM '01), San Jose, California.

Bar-Ilan, J. (2004). The use of web search engines in information science research, In: Cronin, B. (ed.), *Annual Review of Information Science and Technology* 38, Medford, NJ: Information Today Inc, pp. 231-288.

Björneborn, L. & Ingwersen, P. (2001). Perspectives of webometrics. Scientometrics, 50(1), 65-82.

Björneborn, L. (2001). Necessary data filtering and editing in webometric link structure analysis. Royal School of Library and Information Science.

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.

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext*. San Francisco, CA: Morgan Kaufmann.

Cronin, B., Snyder, H.W., Rosenbaum, H., Martinson, A. & Callahan, E. (1998). Invoked on the Web. *Journal of the American Society for Information Science*, 49(14), 1319-1328.

Davenport, E. & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In: Cronin, B. & Atkins, H. B. (eds.).The web of knowledge: a festschrift in honor of Eugene Garfield. Metford, NJ: Information Today Inc. ASIS Monograph Series, 517-534.

Etzkowitz, H. & Leydesdorff, L. (1997). Universities and the global knowledge economy: A triple helix of university-industry-government relations. London, UK: Cassell Academic.

Geisler, E. (2000). The metrics of science and technology. London, UK: Quorum Books.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). The New Production of Knowledge. London, UK: Sage.

Gibbons, M. (1999). Science's new social contract with society. Nature 402, C81-C84.

Google (2003). Google Web APIs. Available: http://www.google.com/apis/, retrieved 24 October, 2003.

Gross, A.G., Harmon, J.E., & Reidy, M.S. (2002). *Communicating Science: The Scientific Article from the 17th Century to the Present*. Oxford: Oxford University Press Inc.

Harter, S. P. & Ford, C. E. (2000). Web-based analyses of e-journal impact: approaches, problems, and issues. *Journal of the American Society for Information Science*, 51(13), 1159-1176.

Hyland, K. (2000), Disciplinary discourses: social interactions in academic writing, Harlow: Longman.

Hyland, K. (2003). Self-citation and self-reference: credibility and promotion in academic publication. Journal of the American Society for Information Science and Technology, 54(3), 251-259.

Ingwersen, P. (1998). The calculation of Web Impact Factors. Journal of Documentation, 54(2), 236-243.

ISI (2003). ISI Web of Science. Available: http://www.isinet.com/isi/products/citation/wos/, accessed 9 October, 2003.

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.

Krippendorff, K. (1980). *Content Analysis; An Introduction to its Methodology*. Beverly Hills CA: Sage.

Landes, W.M. & Posner, R.A. (2000). Citations, age, fame, and the Web. Journal of Legal Studies, 29(319), 329-341.

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the web. Nature, 400, 107-109.

Leydesdorff, L. & Curran, M., (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy, *Cybermetrics*, 4. Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

Liu, C., & Arnett, K.P. (2000). Exploring the factors associated with Web site success in the context of electronic commerce. Information & Management, 38(1), 23-33.

Liu, B., Ma, Y. & Yu, P. (2001). Discovering unexpected information from your competitors' web sites. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, Ca: ACM Press, pp. 144 - 153.

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

Meyer, M. (2003). Academic patents as an indicator of useful research? A new approach to measure academic inventiveness. Research Evaluation, 12 (1), 17-28.

Moed, H. (2002). The impact-factors debate: the ISI's uses and limits, Nature, 415, 731-732.

Moncada, P., Rojo, J., Bellido, F., Fiore, F., & Tübke, A. (2003). Early identification and marketing of innovative technologies: A case study of RTD result valorization at the Joint Research Centre, Technovation, 23, 655-667.

New Scientist (2003). Tech transfer in Germany. New Scientist, 2391, 48-51.

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

Perry M., & Bodkin, C.D. (2002). Fortune 500 manufacturer web sites - Innovative marketing strategies or cyberbrochures? International Marketing Management, 31(2), 133-144.

Posner, R. A. (2001). Public intellectuals: A study of decline. Cambridge, MA: Harvard University Press.

Price, D.J. de Solla. (1963). *Little Science, Big Science*. New York: Columbia University Press.

Rubenstein, A. & Geisler, E. (1991). Evaluating the outputs and impacts of R&D/innovation, International Journal of Technology Management, 5(1), 181-204.

Schmoch, U. (2003). Service marks as novel innovation indicator. Research Evaluation, 12(2), 149-156.

Sloan, B. (2001), Personal Citation Index: Exploring the impact of selected papers, Available: http://www.lis.uiuc.edu/~b-sloan/pci2.html. Accessed 18 June, 2002.

Smith, A.G. (1999). A tale of two web spaces: comparing sites using Web Impact Factors, *Journal of Documentation*, 55(5), 577-592.

Stewart, K.J. (2003). Trust transfer on the World Wide Web. Organization Science, 14(1), 5-17.

Tashakkori, A. & Teddlie, C. (1998). Mixed methodology. London: Sage.

Tijssen, R.J.W. (2003). Scoreboards of research excellence. Research Evaluation, 12(2), 91-103.

Thelwall, M., Vaughan, L. & Björneborn, L. (2005, to appear). Webometrics. In: Cronin, B. (ed.), *Annual Review of Information Science and Technology* 39, Medford, NJ: Information Today Inc.

Thelwall, M. (2002a). Research dissemination and invocation on the Web, *Online Information Review,* 26(6), 413-420.

Thelwall, M. (2002b). The top 100 linked pages on UK university Web sites: high inlink counts are not usually directly associated with quality scholarly content, *Journal of Information Science*, 28(6), 485-493.

Thelwall, M. (2003). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation, Internet Research, 8(3), paper no. 151. Available: http://informationr.net/ir/8-3/paper151.html.

Thelwall, M. (2003b, to appear). Weak benchmarking indicators for formative and semi-evaluative assessment of research, *Research Evaluation*.

van Leeuwen, T.N., Moed, H.F., Tijssen, R.J.W., Visser, M.S., van Raan, A.F.J. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. Scientometrics 51(1), 335-346.

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 analysis of 10 science-intensive technology domains. *Scientometrics*, 57(3), 355-367.

van Raan, A.F.J. (2000). The Pandora's box of citation analysis: Measuring scientific excellence – the last evil? In: Cronin, B. & Atkins, H.B. (Eds.). The Web of knowledge: a festschrift in honor of Eugene Garfield. Metford, NJ: Information Today Inc. ASIS Monograph Series, 301-319.

Vann, K., & Bowker, G. C. (2001). Instrumentalizing the truth of practice, Social Epistimology, 15(3), 247-262.

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), 1313-1324.

Vaughan, L. & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology*, 54(1), 29-38.

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.

Zipf. G. (1932). Selective Studies and the Principle of Relative Frequency in Language. Harvard University Press, Cambridge, MA.

Appendix: Classification Scheme for Web Pages Including the Names of Journals

What is cited?

1 Whole journal
   a) The journal name is given, but never the specific details of any article, e.g. title, authors or page numbers.
2 Individual issue of journal
   a) The journal name is given and a volume number, probably with some kind of description of the volume contents.
3 A specific article in the journal
   a) The specific details of an article, are given, enough to identify it. E.g. title, or authors or page numbers.
4 Other
5 MISTAKE – the journal is not in the page [Ignore all other questions if this is true]

Where is it cited? What type of document contains the citation?

1 The personal page of a person or a list of publications of an individual, not a student.
   a) Gives the person's name and states that it is their personal page
   b) Contains the publications of only one person (some may be joint articles) unless the page is in a database site with a page for every author.
2 The personal page of an individual (undergraduate) student.
   a) Appears to be a personal page, owner states they are a student, or it appears from the context that they are a student.
3 A list of publications of a team of people
   a) The page states that it is a list of publications from a team.
   b) Contains the publications of more than one person and is not a list of publications with one common author.
4 Online journal article or book, or reference list from a book or article
5 Course/class/module page.
   a) Gives the course/class/module title and states that it is a page for it
   b) Is a class reading list.
6 A list of articles in any other context.
   a) At least three articles but not fitting any description above.
7 List of journals [not in a library]
   a) At least three journal names are given in a list form in the page but not details of any individual article [but not in a library site].
8 List of journals in a library [as above but in a library]
   a) It is clear from the URL that it is in a library. E.g. it contains "library" or "lib"
   b) It states in the page that it was created by, or is owned by, a library.
   c) It is clear from the page contents that it is in a library.
9 Other
10 Unknown

Why: what is the relationship between the citer and cited?

1 The author of the article cites it (includes groups where one person is the author)
   a) The page claims that its author wrote the cited article.
   b) The article contains the publications of only one person (some may be joint articles) unless the page is in a database site with a page for every author.
2 The editor of the journal names the journal or lists its articles
   a) The page claims that its author is the journal editor or editorial board member and does not only cite articles written by the page author.
3 Other
4 Unknown

Who owns the citing page?

1 University
   a) The page is from an academic domain, with "edu" or "ac" in the domain name
   b) It says in the page that it is part of a university site.
   c) It is clear from the page that it is from a university

2 Children's School
   a) It says in the page that it is part of a school.
   b) It is clear from the URL that it is part of a school.
3 Commercial company
   a) It says in the page that it is part of a company, or contains its logo.
   b) It is clear from the URL that it is a commercial site, e.g. contains ".co.uk".
   c) It seems to be in a company site, perhaps including a ".com" domain name.
4 Government organisation
5 Other
6 Unknown

Application: Is there any evidence from the page that a commercial company is using the research?

1 yes
   a) It is a commercial company page and the research is mentioned in context with its activities.
2 no
3 Unknown

Link: Is there a link in the page to the journal/journal article?

1 To the journal Web site
   a) There is a hyperlink over the journal name or near it and the URL of the hyperlink points to the journal web site (can check by following it).
2 Direct link to an article
   a) There is a hyperlink over the article name or near it and the URL of the hyperlink points to a copy of the article (can check by following it).
3 Link to some other source of information about the journal
   a) There is a hyperlink over the article name or near it and the URL of the hyperlink points to another type of page containing information about the article (can check by following it).
4 Link to password-protected site
   a) Following the link leads to a request for a username/password to access the page.
5 Link to publisher's site for the journal
   a) Following the link leads to a page about the journal in a big publisher's site. E.g. Elsevier, Wiley, Emerald, Springer, Sage.
6 No link
7 Unknown