An Automatic Method for Extracting Citations from Google Books[1]

Kayvan Kousha
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK E-mail: k.kousha@wlv.ac.uk

Mike Thelwall
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK. E-mail: m.thelwall@wlv.ac.uk

Recent studies have shown that counting citations from books can help scholarly impact assessment and that Google Books (GB) is a useful source of such citation counts, despite its lack of a public citation index. Searching GB for citations produces approximate matches, however, and so its raw results need time-consuming human filtering. In response, this article introduces a method to automatically remove false and irrelevant matches from GB citation searches in addition to introducing refinements to a previous GB manual citation extraction method. The method was evaluated by manual checking of sampled GB results and comparing citations to about 14,500 monographs in the Thomson Reuters Book Citation Index (BKCI) against automatically extracted citations from GB across 24 subject areas. GB citations were 103% to 137% as numerous as BKCI citations in the humanities, except for tourism (72%) and linguistics (91%), 46% to 85% in social sciences, but only 8% to 53% in the sciences. In all cases, however, GB found substantially more citing books than did BKCI, with BKCI's results coming predominantly from journal articles. Moderate correlations between the GB and BKCI citation counts in social sciences and humanities, with most BKCI results coming from journal articles rather than books, suggests that they could measure the different aspects of impact, however.

Introduction

Books are major scholarly outputs in many social sciences and humanities disciplines and are therefore important for research evaluation (e.g., Moed, 2005; Nederhof, 2006; Huang & Chang, 2008). For instance, about a third of the submissions in social sciences and humanities fields to the 2008 U.K. Research Assessment Exercise (RAE) were books in comparison to about 1% in the sciences (Kousha, Thelwall & Rezaie, 2011). Moreover, counting citations from books rather than journal articles can give different results when benchmarking authors (Cronin, Snyder & Atkins, 1997) and countries (Archambault et al., 2006) in the social sciences and humanities. This shows that citations *from* books are an important source of impact evidence that cannot be replaced by citations from journal articles. The lack of a comprehensive index for the bibliographic references of books is therefore an issue for bibliometric monitoring of research in book-based disciplines. Almost two decades ago, this led to a call to include citations from books in academic citation databases (Garfield, 1996). Nevertheless, most previous quantitative investigations into the impact of book-based scholarship have counted citations from journal articles indexed in the commercial citation databases (Web of Science and Scopus) (e.g., Glänzel & Schoepflin, 1999; Butler & Visser, 2006; Bar-Ilan, 2010; Hammarfelt, 2011) rather than citations from other books, although some studies have manually extracted cited references from selected monographs for bibliometric analysis (e.g., Cullars, 1998; Krampen, Becker, Wahner & Montada, 2007). There have also been initiatives to use non-citation metrics for usage assessment of books, such as counting library holdings ("libcitations") (White, Boell, Yu et al., 2009) and using library loan statistics (Cabezas-Clavijo et al., 2013).

Several attempts have been made to extract citations from academic books on a large scale for citation analysis or citation searching. In 2011 Thomson Reuters introduced the Book Citation Index, a set of citations from selected academic books and book chapters that could be added to the journal citations in the Web of Science (WoS). Whilst this is a valuable

[1] This is a preprint of an article to be published in the Journal of the American Society for Information Science and Technology

new source of book-based citations, the current (2013) version of BKCI has partial coverage of English language academic books from selected publishers and therefore should cautiously be used for evaluative purposes (Gorraiz, Purnell & Glänzel, 2013; Torres-Salinas et al., 2012 and 2013). It is also possible to use carefully constructed queries to identify citations in books using the GB search interface (Kousha & Thelwall, 2009; Kousha, Thelwall & Rezaie, 2011). Nevertheless, the GB citation searching needs extensive human labour to manually locate accurate citations from the GB search results, which typically include substantial numbers of approximate matches in addition to any correct matches.

The current study introduces, applies and assesses a new automatic method to extract citations from GB to eliminate the human labour needed for the previously used method (in Kousha & Thelwall, 2009). Citations to about 14,500 BKCI-indexed books were automatically extracted from GB using this method and compared with BKCI citations in order to identify the strengths and weaknesses of the two data sources.

Background

Google Books

GB (http://book.google.com) is a large collection of digitised academic and non-academic books that is constantly growing by reading or scanning books from the libraries and publishers (see also: Vincent, 2007). In addition to citations, GB is interesting for the legal issues involved in the online availability of books (Travis, 2010; Fulda, 2012) and for its application in library services (Jackson, 2008; Leonardo, 2012).

GB apparently covers about 30 million volumes (Darnton, 2013), although exact details of the books included seems to be a commercial secret. GB clearly indexes a substantial fraction of the world's books, however. Out of 401 randomly selected books from WorldCat (which claims to be the world's largest library catalog) in different languages, metadata for 84% (336) were found in GB (Chen, 2012), suggesting that GB is quite comprehensive on an international scale. A minority of these books had full-text views (8%), previews (16%, limited views of any pages) and snippets (13%, limited views of a few sentences around the search terms) (Chen, 2012). Hence, it seems that over a third (37%) of all GB books and about a fifth (21%) of its WorldCat books have the full-text search capability in GB that is needed for citation searching. The same approach was applied to a random sample of 1,500 Hawaiian and Pacific books from a university library collection in the United States, with similar results. About 80% of the sampled books were found in GB and a third (32%) of the books were fully searchable (Weiss & James, 2013). GB coverage of 87 core medical textbooks was found to be much higher, however: the metadata of only three titles was not found and about two-thirds (64%) were fully searchable in GB (Johnson & Lottes, 2009).

GB includes some errors that were presumably generated by the automatic scanning process. James and Weiss (2012) examined metadata (e.g., author, title, publisher, publication year) from 400 randomly selected scanned texts, finding that 36% contained metadata errors. Of these errors, 41% were related to publishers' names, 24% to authors' names, 20% to publication dates and 15% to titles. These metadata errors should have little impact on the methods used to extract citations from GB, however, since these use the scanned text rather than metadata. In contrast, out of 2,500 pages from 50 randomly selected books, less than 1% had legibility errors (James, 2010) so the scanned text appears to be much more reliable than the metadata.

Google Books Citations

Although not a citation index, the GB full-text search can be used to locate citations in the text of digitised books. Two investigations have previously used online GB searches for manual citation extraction. A comparison of citations from online GB searches with WoS citations to over 3,500 journal articles in ten fields found that GB citations were 31%-212% as numerous as WoS citations in the social sciences and humanities, but only 3%-5% in the sciences, except for computing (46%) (Kousha & Thelwall, 2009). There were significant Spearman correlations between GB and WoS citation counts in the selected subject areas and this relationship was higher in social sciences and humanities than in the sciences (except for computing). The study concluded that GB citation search is valuable for research evaluation in book-based fields, although very time-consuming for large scale assessments. A follow-up study compared citations from GB searches with Scopus cited reference searches to books rather than journal articles (Kousha, Thelwall & Rezaie, 2011). Using 1,000 books submitted to the 2008 U.K. Research Assessment Exercise in seven book-oriented subject categories, citations from GB to other books were found to be 1.4 times more numerous than citations from Scopus articles to books. This suggests that GB citations can give evidence of research impact to assist peer-review in book-oriented fields.

The Thomson Reuters Book Citation Index

Until 2011, the lack of a substantial coverage of books in WoS and related Thomson Reuters citation databases caused problems for research performance monitoring in the social sciences and humanities (e.g., Hicks, 1999; Nederhof, 2006). For instance, a large-scale study of references from social sciences and humanities articles published 1981-2000 showed that the proportion of citations to journal articles from Thomson ISI (Now Thomson Reuters) indexed publications was almost half that for natural sciences and engineering (45% vs. 86% respectively), indicating that in the social sciences and humanities, non-serials and monographs are the majority sources of evidence, even for journal articles (Larivière, Archambault, Gingras & Vignola-Gagné, 2006).

To include some citations from monographs, Thomson Reuters launched the BKCI in 2011 through the WoS interface, starting with books published in 2005. There were initially about 40,000 books and about 10,000 new titles were added each year, mostly from the social sciences, arts and humanities (60%), with the remainder covering science and medicine (The Book Citation Index, http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/, July 2013). The selection process is mainly restricted to English scholarly books that "present fully referenced articles of original research, or reviews of the literature". BKCI also includes textbooks for graduate or advanced research and translations of non-English works (Testa, 2011).

It is important to discriminate between monographs and edited books because they have different citation characteristics. For example, book chapters can be highly cited and have been detectable in WoS since 2005 - in biochemistry in particular (Leydesdorff & Felt, 2012). In contrast, although books in BKCI tend to have many references, they tend to be relatively less cited (Leydesdorff & Felt, 2012).

In book-based subjects, the publisher of a book can be as important as the journal publishing an article in other areas. In consequence, 'Book Publishers Citation Reports' have been proposed based on BKCI data in analogy with the 'Journal Citation Reports' (Torres-Salinas et al., 2012). Nevertheless, problems of name variations of publishers, over-representation of English-language books, and low representation of many countries with substantial social sciences and humanities publishing (e.g., Italy, France, Germany) are major obstacles for the development of such an indicator (Torres-Salinas et al., 2012). Moreover, since most BKCI-indexed book chapters are from a few publishers (e.g., Springer, Routledge,

Palgrave, Nova Science) across all subject areas, BKCI seems to be unbalanced and missing major publishers in some fields (Torres-Salinas et al., 2013; Gorraiz, Purnell & Glänzel, 2013). This issue not only undermines the Book Publishers Citation Reports idea but also the use of BKCI for citation analysis (Torres-Salinas et al., 2013; Gorraiz, Purnell & Glänzel 2013).

Research questions

The main aim of this study is to assess whether a new GB automatic citation extraction method, described in the methods section, could be useful for the citation impact assessment of academic books in science, social science and the humanities. If so, then this would make large-scale research evaluation possible based upon books. The following research questions guide this investigation.

1. Can GB automatic citation extraction give sufficient results to be a viable alternative to BKCI for the impact assessment of books?
2. How do disciplinary differences and the time from the publication of a book influence the answers to the above question?

Methods

To address the above research questions, an automatic method was developed to: (a) identify potential mentions of books in GB through queries submitted to its Application Programming Interface (API); (b) apply matching and filtering techniques to identify correct citations and remove incorrect approximate matches; and (c) to remove unwanted types of matches (e.g., advertisements, book reviews and bibliographies) from the results. The method was then evaluated and applied to compare GB citations against BKCI citations to books from 24 science, medicine, social sciences, and arts and humanities subject areas.

Research Population

Bibliographic information was extracted from BKCI for 14,487 monographs published during 2005-2010 in 24 fields. The years 2005-2010 were selected to give books at least two years to receive citations from other books and also to analyse the impact of time on the results. The 24 fields were selected to represent a range of different subject areas within each broad category. A generic alphabetical query[2] was used to retrieve all book records from the *Book Citation Index-Social Sciences & Humanities* (BKCI-SSH) and the *Book Citation Index-Science* (BKCI-S), limiting the results to '*Books*' (excluding articles, editorial materials, biographical items, and reviews) within the selected subject areas. Edited books or volume series were then removed from the BKCI outputs to limit the data set to monographs through excluding records with '*Book Editor(s)*' (the *BE* field in the BKCI output) and titles either ending with '*edition'* (e.g., *Laser Material Processing, 4th Edition*) or with different volumes (e.g., *Mechanical Systems, Classical Models, Vol II*). Monographs alone were selected because the citations to individual book chapters and volume series are not included in the count of citations to whole books in BKCI, so the BKCI citation counts for edited volumes could be significant underestimates in many cases. In contrast, GB citation searches can return citations to whole edited books and their chapters, so GB and BKCI do not give comparable results for edited works and volume series. Related subject areas were merged based on WoS categories (see Table 1) to have a large data set for analysis and to avoid as far as possible

2. The query used in the "publication name" field was: (A* OR B* OR C* OR D* OR E* OR F* OR G* OR H* OR I* OR J* OR K* OR L* OR M* OR N* OR O* OR P* OR Q* OR R* OR S* OR T* OR U* OR V* OR W* OR X* OR Y* OR Z* OR 0* OR 1* OR 2* OR 3* OR 4* OR 5* OR 6* OR 7* OR 8* OR 9*)

using individual books multiple times if they were indexed in two or more subject categories. All BKCI and GB data collection took place during April 2-5, 2013.

Table 1. Sources of books selected from BKCI (2005-2010).

Broad Fields	Subject Area	No. of Books
Social Sciences (4,324)	**Business** (Economics; Management; Finance)	840
	Education (Educational Research; Education, Special)	660
	Psychology (Clinical; Multidisciplinary; Experimental; Applied)	440
	Sociology (Anthropology; Ethnic Studies; Women's Studies; Cultural Studies)	731
	Political Science (International Relations)	838
	Social Sciences (Social Work; Social Issues; Interdisciplinary)	406
	Geography (Urban Studies; Area Studies; Demography)	207
	Information and library Science	202
Arts and Humanities (5,724)	**History**	981
	Law (Criminology; Penology)	309
	Literature (Classics; Poetry; Literary Theory & Criticism)	1,480
	Art (Music; Dance; Theater; Film, Radio & Television)	721
	Philosophy (History & Philosophy of Science)	624
	Religion (Medieval & Renaissance Studies)	868
	Linguistics (Language Studies)	532
	Tourism (Hospitality, Leisure & Sport; Transportation)	209
Sciences and Medicine (4,439)	**Chemistry** (e.g., Organic; Applied; Analytical; Multidisciplinary; Chemistry, inorganic & nuclear; Chemistry, medicinal)	214
	Computer Science (e.g., Software Engineering; Hardware & Architecture; Artificial Intelligence; Automation & Control Systems; Theory & Methods)	751
	Engineering (e.g., Civil; Electrical & Electronic; Mechanical; Materials Science; Telecommunications; Industrial; Environmental; Biomedical; Multidisciplinary)	944
	Environmental Sciences (Ecology; Marine & Freshwater Biology; Parasitology; Biodiversity Conservation; Meteorology & Atmospheric Sciences; Oceanography)	244
	Mathematics (e.g., Applied; Statistics & Probability)	1,207
	Medical Sciences (e.g., General & Internal; Public Health; Surgery; Immunology; Veterinary Sciences; Neurosciences & Neurology; Pharmacology & Pharmacy; Genetics & Heredity; Radiology; Dentistry)	575
	Biotechnology (e.g., Biochemistry & Molecular Biology; Cell Biology; Cell & Tissue Engineering; Applied Microbiology)	96
	Physics (e.g., Applied; Particles & Fields; Optics; Atomic; Condensed Matter)	408
Total (all fields)		**14,487**

Google Books Automatic Citation Extraction

GB allows full-text searching for some digitised books and gives different levels of access including *full view* (free full-text and fully searchable books), *preview* or *snippet view* (fully searchable books but the results are displayed in sample pages or few sentences around search term from pages), and "*no preview*" (non-searchable, non-viewable books). Hence, for books with full-text searching capability, citations can be identified from reference lists, footnotes or the main text (see Example 1: http://cybermetrics.wlv.ac.uk/paperdata/GBExamples.doc).

Identifying Citations with Google Books API Searches

GB supports automatic searching with its API (see: https://developers.google.com/books/docs/v1/getting_started). Code to gather data from the GB API and to implement the automatic filtering was added to the free software *Webometric Analyst* (http://lexiurl.wlv.ac.uk) (see its "Books" tab). The software generates and runs queries to locate GB citations by feeding a list of publications from either WoS or Scopus outputs.

Different GB queries for book citations using the bibliographic information from BKCI were tested in an attempt to create a search strategy with the highest accuracy and coverage of formal citations. Each query tested contained at least one author last name, the publication year, and some words from the book title. For instance, the tests revealed that phrase searches for long book titles often substantially reduced the number of correct search results (lowering recall). In contrast, using short title searches (e.g., the first two words) tended to substantially increase the number of false matches (lowering precision), especially when authors' or editors' last names were very common.

- *Knox "To the Threshold of Power, 1922/33: Origins and Dynamics of the Fascist and National Socialist Dictatorships" 2007* = **14 citations**
- *Knox "To the Threshold of Power, 1922/33: Origins and Dynamics" 2007* = **22 citations**

To investigate this key issue, a random sample of 700 books across different fields with six or more terms in their titles was investigated. Two queries were generated for each book, one with a phrase search of the first six terms in the title and one with a phrase search of the full book title. These titles were combined with the last name of the first author or editor and the publication year. Both queries were searched through *Webometric Analyst* at the same time and citation counts were recorded after removing false matches. The number of correct GB citations for queries truncated to six title terms was significantly higher (median 6) than for queries with seven or more terms (median 4) for the same books (Wilcoxon signed ranks test, p=0.000).

One reason for the reduced effectiveness of queries with longer titles in GB is the occurrence of non-alphanumeric characters, such as punctuation, in titles which seems to increase the false citations rate (also reported in: Kousha, Thelwall, & Rezaie, 2011). Meta-information within the BKCI-reported title of a book could also influence the number of matches for long title searches, such as specifications of an edition, volume or version of a book, as the underlined parts of the examples below illustrate.

- Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do, Expanded Edition
- Treatment Approaches for Alcohol and Drug Dependence: An Introductory Guide, 2nd Edition
- Etudiants De L'Exil: Migrations Internationales et Universites Refuges (XVI-XX s.)

In the above cases a citation style might mention the same information in a different way (e.g., "2nd Ed." instead of "2nd Edition") or separately from the title. More fundamentally, it might make sense from an impact assessment perspective to combine the citations to all volumes, editions or versions of a text. There were also errors and missing apostrophes in some titles from BKCI data and this was the main reason for searches without any valid results for long titles.

- Asia's New Mothers: Crafting Gender Roles and Childcare Networks in East amd Southeast Asian Societies **["amd" instead of and]**
- Complex Adaptive Systems: An Introduction to Computational Models of Social Life: An Introduction to Computational Models of Social Life **[repeated subtitle]**
- Schoolhouses, Courthouses, and Statehouses: Solving the Funding-Achievement Puzzle in Americas Public Schools **[missing apostrophe]**

Including publisher names in the queries tended to significantly reduce the coverage of the citation searches because these can be written in multiple different ways, as the following

examples for a single book illustrate. For this reason publisher names and places of publication were not added to the queries.

"Academic Press/Elselvier [sic], Amsterdam"
"Academic, New York"
"New York: Elsevier/Academic Press"
"Academic Press: San Diego, CA"
"Academic Press, London",
"London: Elsevier"

The query format that was eventually chosen combined the last name of the first author or editor, a phrase search for the first six terms in the title (or the full title if it had fewer than six terms) and the publication year.

- *Step 1*: Query GB using the format [*first author or editor last name] "[the six first terms in the title]" [publication year]*.

Previous experiments with GB citation showed that many false matches occurred for books with very general single or two word titles (e.g., "*Doubt*" or "*Music Perception*") and common last author names (e.g., *Smith* or *Jones*). Although only 3% (779) of the books in the current study had either less than three words in their titles, we added the place of publication, as recorded in BKCI, to the queries for these cases to reduce the number of false results.

Removing Incorrect Matches from Google Books API Searches

The next task was to remove search matches that were incorrect in the sense of not mentioning the correct book. For instance, the query *Moed "Citation analysis in research evaluation" 2005* in GB returned 15 citations inside other books as well as many incorrect matches, such as "*Noisy Poems"* and "*Dear Mum, I miss you!*", which do not cite Moed's book (see Example 2: http://cybermetrics.wlv.ac.uk/paperdata/GBExamples.doc). This shows that GB uses approximate matching and so its results must be filtered. This is a change from previous experiments using citation matching with GB, which did not find the same problem.

The GB results include a description field that can be used to assess whether a search match is correct or not. In the online version of GB search this is shown as the snippet of text describing each result. In the case of a correct match, the description tends to contain the citation itself, or at least the part of the citation that contained the query terms (the first six terms of the book title, the author last name and the publication year). Hence, query matches not containing the query terms within the description field were automatically removed.

- *Step 2*: remove query matches that do not contain the query terms within their description field.

Assessing Recall and Precision for GB Automatic Searches

A manual check of 335 randomly sampled results from steps 1 and 2 from all three broad areas gave an overall precision of 91%. Additional manual searches were used for each book in order to identify any obvious cases of missing relevant results and this produced 184 new citations (i.e., the automatic method missed at least 8% of the citations, a recall of up to 92%). The missing relevant results were mainly due to errors in author last names or book titles (see the next section).

Removing Correct Matches but False Citations

A set of rules was devised to remove matches that were technically correct in the sense that they mentioned the right book, but were conceptually false, in that the mentioned book was not formally cited. This extra step excluded about 8.5% of the results from 182,831 initial GB automatic searches of all books (24,140) in the study. The list below (and Example 3: http://cybermetrics.wlv.ac.uk/paperdata/GBExamples.doc) gives more details of the methods and reasons for these exclusions.

- *Step 3a: Search results matching "bibliogr*", "book review*", "abstracts" in their titles were excluded as well as the title "Choice".* Bibliographies, book reviews and abstracting and indexing volumes all contain complete details of books without citing them. In addition, *Choice* contains *American Library Association* book reviews. This step excluded about 3,600 matches (2% of the initial 182,831 GB results, including about 2,150 results from *Choice*).
- *Step 3b: Search results with titles matching the citation were excluded*. Self-mentions of books often occurred in cataloguing records, front pages and back covers. This step excluded about 1,750 matches (1% of the initial GB results).
- *Step 3c: Search result descriptions containing price signs or any one of a set of identified phrases representing publisher advertisements (e.g., "Series Editor:") were excluded*. Many publishers advertise books inside other books. These tended to include book prices (e.g., $, USD, £, GBP) or a few common phrases. This step excluded about 2,650 matches (1.4% of the initial GB results).
- *Step 3d: Search result descriptions containing* ISBN, hardback, *or* paperback *were excluded.* The above steps still left many book lists, often including of the terms *ISBN*, *hardback*, or *paperback*. These terms seem to be rarely used in traditional citations. This step excluded about 3,700 matches (about 2% of the initial GB results).
- *Step 3e: Search result descriptions containing author self-descriptions (e.g., "is professor", "is a professor) were excluded.* Author biographies often mention their previous or in press books in notes on authors and in sections about book chapter contributors. A range of short phrases commonly used in such texts was manually compiled and tested in order to exclude these. This step excluded about 4,100 matches (2.2% of the initial GB results).

Increasing Coverage (Recall) of the Citation Results

As discussed above, the automatic method did not retrieve at least 8% of the possible GB citations. Additional manual checks revealed the fact that BKCI had merged many compound last names (e.g., VanDerWurf or SutherlandAddy instead of Van Der Wurf and Sutherland-Addy), omitted accents (e.g., Duhr instead of Dühr) and omitted apostrophes in titles (womens instead of women's).

To solve the first problem, when building the GB queries from BKCI, author last names were automatically split whenever a lower case letter was followed by an upper case letter. The initial letters "Mc" were ignored as an exception for splitting names in queries (e.g., McNeill). New searches were then conducted to assess whether extra citations could be identified with the revised names. The matching process in Webometric Analyst was also revised to ignore accents on characters in last names and non-alphanumeric characters in titles when matching the initial queries against GB description results. These solutions increased the number of relevant citations by about 3,000 (2% additional relevant results).

Step 1 modification: Split author last names whenever a lower case letter is followed by an upper case letter, except for Mc.

Step 2 modification: remove apostrophes, accents from characters and non-alphanumeric characters before checking queries against search results descriptions.

Results

Non-parametric Mann-Whitney U tests showed that the difference between GB and BKCI citations is statistically significant in arts and humanities ($p = 0.000$) and in sciences and medicine ($p = 0.032$), but not in the social sciences. The overall results show that the total and median GB citations to books in arts and humanities are significantly higher than BKCI citations (GB citation is 118% of BKCI citations), indicating that GB coverage of books is

better for citation analysis in book-based disciplines. In contrast, in sciences and medicine BKCI citations are considerably more numerous than GB citations (BKCI citation is 385% of GB citations) due to many citations from WoS-indexed journal articles (Table 2). In eight social science fields there are more total BKCI citations than GB citations (BKCI citation is 169% of GB citations), the medians of GB and BKCI citations are the same (4) and the difference between the overall distribution of GB and BKCI citations is not statistically significant ($p = 0.350$). In eight sciences and medicine there are 385% more BKCI citations than GB citations but both have equal medians (1). This indicates that the distribution of citations in BKCI is highly skewed in science due to many highly WoS-cited monographs that received relatively few GB book citations. The high values for BKCI is possible because many BKCI citations come from journal and conference papers in WoS, whereas GB citation searches only includes books.

There are significant ($p<0.01$) moderate Spearman correlations between the BKCI and GB citation counts in both the social sciences (0.581) and humanities (0.570). The correlation for science and medicine is significant but much lower (0.263), perhaps because monographs are less important for transmitting scientific research. In all three broad fields correlation between filtered GB citations and BKCI citations is slightly higher than the correlation between the raw GB and BKCI citations (the final column of Table 2). This difference in correlations is evidence that the filtering method improves the quality of the results. This is because if the removed results are predominantly incorrect then they will have a correlation of close to zero with BKCI citations, because there is no reason for them to be related, so removing predominantly incorrect results would increase the correlation, whereas removing predominantly correct results would not be likely to affect the correlation much.

Table 2. Citations from GB and BKCI and correlations between them in three broad areas.

Broad Fields	Books	GB citations incl. false matches	GB citations (filtered)			Thomson Reuters BKCI			Correl.: GB and BKCI (*raw GB and BKCI*)
			Citations	Median (mean)	% of BKCI cites	Citations	Median (mean)	% of GB cites	
Social Sciences	4,324	159,457	37,948	4 (8.8)	59%	64,213	4 (14.8)	169%	0.581** *(0.463**)*
Arts and Humanities	5,724	242,600	54,086	5 (9.4)	118%	45,832	4 (8.0)	85%	0.570** (*0.436***)
Sciences and Medicine	4,439	113,647	17,410	1 (3.9)	26%	66,957	1 (15.1)	385%	0.263** (*0.221***)
Total	**14,487**	**515,704**	**109,444**	**3 (7.4)**	**62%**	**177,002**	**3 (11.9)**	**162%**	**0.483**** (*0.419***)

***Significantly different from 0 at p=0.001.*

Disciplinary Differences

There is a wide variation between individual disciplines in terms of the ratio of GB to BKCI citations, even within the same broad area (Table 3). GB citations were 72%-137% as numerous as BKCI citations in the humanities, 46%-85% in the social sciences and 8%-53% in the sciences. In conventional arts and humanities book-based fields, such as law, literature, history, philosophy and religion, GB citations are more numerous than BKCI citations and useful in research assessment, but are less plentiful in linguistics and tourism, suggesting the significance of journal articles in these fields. In all social science fields there were more BKCI citations than GB citations. The GB median was higher in two cases (education and political science) and lower in three (geography, information science, interdisciplinary social sciences) suggesting that both books and journal articles are commonly used for research communication in the social sciences. Unsurprisingly, in journal-based fields, such as chemistry, physics and engineering, BKCI citations were many times more numerous than GB

citations respectively, indicating the majority of citations to monographs coming from journal articles rather than books.

Table 3. Citations from GB and BKCI and correlations between them in each studied discipline.

Broad Fields	Disciplines	No. of books	GB Results incl. false matches	GB Citations (filtered)			BKCI			Correl.: GB and BKCI
				No. of citations	Median (mean)	% of BKCI cites	No. of citations	Median (mean)	% of GB cites	
Social Sciences	Business	840	26,859	5,215	2 (6.2)	48%	10,856	2 (12.9)	208%	0.570**
	Education	660	19,902	4,302	3 (6.5)	85%	5,063	2 (7.7)	118%	0.503**
	Psychology	440	15,506	3,838	4 (8.7)	48%	8,026	4 (18.2)	209%	0.547**
	Sociology	731	31,232	7,943	6 (10.9)	56%	14,103	6 (19.3)	178%	0.581**
	Political Sci.	838	33,612	8,917	6 (10.6)	74%	12,070	5 (14.4)	135%	0.636**
	Social Sci.	406	16,718	4,380	5 (10.8)	46%	9,510	6 (23.4)	217%	0.559**
	Geography	207	9,346	2,177	5 (10.5)	72%	3,027	7 (14.6)	139%	0.567**
	Inform. Sci.	202	6,282	1,176	2 (5.8)	75%	1,558	3 (7.7)	132%	0.490**
Arts and Human.	History	981	46,904	11,659	8 (11.9)	122%	9,525	6 (9.7)	82%	0.621**
	Law	309	10,646	2,698	5 (8.7)	137%	1,968	1 (6.4)	73%	0.525**
	Literature	1,480	59,277	11,062	4 (7.5)	133%	8,308	3 (5.6)	75%	0.532**
	Art	721	28,766	5,726	4 (7.9)	131%	4,370	3 (6.1)	76%	0.594**
	Philosophy	624	29,717	5,894	4 (9.4)	103%	5,734	3 (9.2)	97%	0.530**
	Religion	868	38,012	9,619	6 (11.1)	130%	7,425	3 (8.6)	77%	0.574**
	Linguistics	532	23,830	6,307	6.5 (11.9)	91%	6,940	6 (13)	110%	0.537**
	Tourism	209	5,448	1,121	2 (5.4)	72%	1,562	3 (7.5)	139%	0.628**
Sciences and Medicine	Chemistry	214	4,337	625	1 (2.9)	8%	7,439	3 (34.8)	1190%	0.308**
	Computer Sci.	751	22,529	3,805	2 (5.1)	36%	10,487	1 (14)	276%	0.236**
	Engineering	944	21,740	2,785	1 (3)	22%	12,650	1.5 (13.4)	454%	0274**
	Environ. Sci.	244	6,412	985	1 (4)	26%	3,832	1 (15.7)	389%	0.495**
	Mathematics	1,207	35,434	5,980	2 (5)	31%	19,343	1 (16)	323%	0.190**
	Medical Sci.	575	12,809	1,956	1 (3.4)	43%	4,553	1 (7.9)	233%	0.431**
	Biotech.	96	1,455	173	0 (1.8)	53%	329	0 (3.4)	190%	0.151**

	Physics	408	8,931	1,101	1 (2.7)	13%	8,324	2 (20.4)	756%	0.263**
Total		**14,487**	**515,704**	**109,444**	**3 (7.4)**	**62**%	**177,002**	**3 (11.9)**	**162** %	**0.483****

GB and BKCI Citations over Time

Both GB and BKCI citation medians seem to increase over time in the long term, so that even the median citations for a seven year time period (2005-2012) are greater than the medians for a six year time period for both GB and BKCI (Table 4). This suggests that long time periods are useful for the impact assessment of monographs in both GB and BKCI.

Table 4. Median and total citations for BKCI books published 2005-2010 from GB and BKCI.

Median/ Total citations	2005		2006		2007		2008		2009		2010	
	GB	BKCI	GB	BKCI	GB	BKCI	GB	BKCI	GB	BKCI	GB	BKCI
Social Sciences	13 13,380	10 25,826	8 6,986	6 9,776	6 6,309	6 11,298	4 5,045	4 7,141	2 4,176	2 6,992	1 2,052	2 3,180
Arts and Humanities	13 17,709	7 14,856	12 12,123	7 8,869	7 10,309	5 9,293	4 7,531	3 6,115	2 4,597	2 4,403	1 1,817	1 2,296
Sciences and Medicine	5 3,850	16 23,689	4 2,455	5 12,914	3 3,262	6 11,520	1 3,279	0 6,184	0 2,925	0 6,378	0 1,639	1 6,272
Total	11 34,939	9 64,371	9 21,564	6 31,559	5 19,880	6 32,111	3 15,854	2 19,440	2 11,696	1 17,773	1 5,511	1 11,748

Discussion

The automatic GB citation extraction method has some limitations. Although the testing described in the methods section seems to give a high overall accuracy and coverage for the automatic GB citation searches (over 90%), the filtering is all based upon heuristics and so it is possible that the results will be poor for some individual books. For instance, rules to filter out results with ISBNs or prices in the GB search results description field will not work in rare cases when they occur in formal cited references, such as in the title of a book or erroneously added to a reference. Another limitation is that the method used has variations for different books. For example books with short titles had extra bibliographic information added to their queries (see methods), and so their queries are likely to have lower recall. Query problems may also affect some areas more than others. Scholars in some humanities fields (e.g., literary studies) may use references in the text such as footnotes, endnotes, and in the main text more frequently for in-depth arguments (Hammarfelt, 2011) which seem to be more difficult to capture in GB. Moreover, the results include lists of books that are not cited but which are 'further readings', 'additional readings', 'key readings', especially in textbooks. These were kept in the results as they seem to be indicators of some kind of intellectual impact, but perhaps their value is less than that of citations. Finally, we restricted the data set to monographs and future research could analyse the difference between citations to book chapters and edited series and monographs, and perhaps also investigate the impact of disciplinary differences.

An important consideration when comparing the BKCI and GB results is that BKCI integrates a large number of non-books within its citation counts (Table 5). In social sciences and humanities about 79% and in sciences about 92% of BKCI citations came from articles indexed by WoS databases, in comparison to 16% and 5% for book, respectively (both books and book chapters). Thus, it seems that GB reports substantially more citations from books than does BKCI, even in the sciences. Since the GB and BKCI citations are mainly from

different publication types and the correlations between the GB and BKCI citation counts across subject areas are only moderate, this is evidence that they may reflect different types of impact.

Table 5. Types of BKCI citing sources to books in social sciences and humanities and sciences

BKCI database	No. of books 2005-2009	Types of citing sources to books as reported in BKCI				Total citing sources
		Articles, reviews and proceeding papers (WoS citations)	Book Chapters	Books	Other (e.g., letters, editorial)	
Social Sciences & Humanities	15,496	112,404 (78.9%)	15,091 (10.6%)	8,196 (5.8%)	6,788 (4.8%)	142,479 (100%)
Science	8,233	98,065 (92.3%)	4525 (4.3%)	870 (0.8%)	2788 (2.6%)	106,248 (100%)

Citations to the top 20 highly cited books from GB without any BKCI citations in three broad fields were manually checked in order to assess whether books that could be uniquely identified as important by GB were genuinely important or whether they were anomalies caused by errors in the automatic GB search process. This sample was taken from the full original data set, including edited books. The overall precision was about 93% (Table 6), confirming that the books without BKCI citations were high impact publications and that the GB results were mostly correct.

Table 6. The accuracy and coverage of GB automatic searches for top 20 GB highly cited books without BKCI citations in three broad areas.

Broad Fields	Highly cited GB books without BKCI citations (Max- Min of GB cites)	Precision	Estimated recall	Total automatic GB search citations (GB initial hits)	Relevant GB results after manual checking	False matches retrieved by automatic search	Missing relevant results from automatic search
Social Sciences	20 (42-29)	94.8%	96.0%	707 (1,903)	698	37 (5.2%)	28 (4.0%)
Arts and Humanities	20 (47-35)	93.2%	97.2%	813 (1,986)	777	55 (6.8%)	22 (2.8%)
Sciences and Medicine	20 (53-26)	90.1%	94.9%	636 (1,876)	602	63 (9.9%)	31 (5.1%)
Total	**60**	**92.8%**	**96.1%**	2,156 (5,765)	**2,077**	**155 (7.2%)**	**81 (3.9%)**

Conclusions

In answer to the first research question, the new automatic GB citation extraction method seems to give sufficient results for it to be useful in research assessment. Moreover, its value is corroborated by its significant correlation with BKCI citations, and with this correlation being higher for the filtered results than for the unfiltered results. Within the arts and humanities, it has a clear advantage over BKCI in terms of the total number of citations found. This is not true for the social sciences and science due to the inclusion of journal articles in the BKCI results. For the social sciences, the BKCI and GB results are very broadly similar in size so for social sciences research assessment it would be reasonable to combine GB results

with WoS results without the books from the BKCI, but it would be better to include BKCI in order to get the widest possible range of sources of impact. For the sciences, it does not seem worth gathering citations through GB because of the lower regard for books amongst scientists and the lower proportion of GB citations compared to BKCI citations for science and medicine.

In answer to the second research question, there were substantial disciplinary differences between GB and BKCI citations across and within the three broad areas. In book-based disciplines, such as law, history, literature, art, philosophy and religion, GB citations are clearly more numerous than BKCI citations. In contrast, in sciences such as physics, chemistry, computing, engineering and mathematics BKCI citations are generally much higher than GB citations due to integrating WoS citations within BKCI.

There are several additional general advantages with using GB automatic citation searches, including the free nature of the GB API and its huge coverage of books, as verified above. With this new method it is now also possible to conduct large-scale impact assessment studies based on books with little human labour. Although GB citations could be a useful indicator to assist the peer-review process in book-oriented fields, fully automatic searches should cautiously be used for individual assessment of academics due to the possibility of significant numbers of false matches for individual books (see the discussion) and so academics should have the right to check their results if they believe them to be incorrect.

Acknowledgments

This paper is supported by ACUMEN (Academic Careers Understood through Measurement and Norms) project, grant agreement number 266632, under the Seventh Framework Program of the European Union. Thanks to Mahshid Abdoli for manual checking the GB citations. An early and partial version of the automated method described in this article was used but not tested in a previous paper (Abdullah & Thelwall, in press).

References

Abdullah, A. & Thelwall, M. (in press). Can the impact of non-Western academic books be measured? An investigation of Google Books and Google Scholar for Malaysia. *Journal of the American Society for Information Science and Technology*.

Archambault, E., Vignola-Gagne, E., Cote, G., Lariviere, V., & Gingras,Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, 68(3), 329–342.

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, 82(3), 495-506.

Butler, L. & Visser, M. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66(2), 327–343.

Cabezas-Clavijo, A., Robinson-García, N., Torres-Salinas, D., Jiménez-Contreras, E., Mikulka, T., Gumpenberger, C., Wemisch, A. & Gorraiz, J. (2013). Most borrowed is most cited? Library loan statistics as a proxy for monograph selection in citation indexes. In*: Proceedings of 14th International Conference of the International Society for Scientometrics and Informetrics*, Vienna, Austria, Vol. 2, pp. 1237-1252. Retrieved October 02, 2013, from http://arxiv.org/ftp/arxiv/papers/1305/1305.1488.pdf.

Chen, X. (2012). Google Books and WorldCat: A comparison of their content. *Online Information Review*, 36(4), 507-516.

Cronin, B., Snyder, H. & Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: a study of sociology. *Journal of Documentation*, 53(3), 263-273.

Cullars, J. (1998). Citation characteristics of English-language monographs in philosophy. *Library & Information Science Research*, 20(1), 41–68.

Darnton, R. (2013). *The national digital public library is launched*. New York Review of Books, 60(7), Retrieved July 02, 2013, from http://www.nybooks.com/articles/archives/2013/apr/25/national-digital-public-library-launched/

Fulda, J. (2012). Google Books and other internet mischief. *Journal of Information Ethics*, 21(2), 104-109.

Garfield, E. (1996). Citation indexes for retrieval and research evaluation. Consensus Conference on the Theory and Practice of Research Assessment, Capri, July 09, 2013, from http://www.garfield.library.upenn.edu/papers/ciretreseval-capri.html

Glänzel, W. & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management*, 35(1), 31−44.

Gorraiz, J., Purnell, P. J., & Glänzel, W. (2013). Opportunities for and limitations of the book citation index. *Journal of the American Society for Information Science and Technology*, 64(7), 1388-1398.

Hammarfelt, B. (2011). Interdisciplinarity and the intellectual base of literature studies: Citation analysis of highly cited monographs. *Scientometrics*, 86(3), 705-725.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2), 193−215.

Huang, M. & Chang, Y. (2008). Characteristics of research output in social sciences and humanities: from a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, 59(11), 1819–1828.

Jackson, M. (2008). Using metadata to discover the buried treasure in Google Book search. *Journal of Library Administration*, 47(1-2), 165-173.

James, R. (2010). An assessment of the legibility of Google Books. *Journal of Access Services*, 7(4), 223-228.

James, R. & Weiss, A. (2012). An assessment of Google Books' metadata. *Journal of Library Metadata*, 12(1), 15-22.

Johnson, E. & Lottes, J. (2009). Google Book Search coverage of core clinical textbooks. In: Positioning the Profession: the Tenth International Congress on Medical Librarianship, Brisbane, Australia, August 31-September 4, 2009. pp. 1-8, Retrieved July 07, 2013, from http://espace.library.uq.edu.au/eserv/UQ:179767/n6_4_Fri_Johnson_71.pdf

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

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

Krampen, G., Becker, R., Wahner, U. & Montada, L. (2007). On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. *Scientometrics*, 71(2), 191–202.

Larivière, V., Archambault, É., Gingras, Y., & Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8), 997-1004.

Leonardo, D. (2012). Google Books: Primary sources in the public domain. *Collection Building*, 31(3), 103-107.

Leydesdorff, L. & Felt, U. (2012). Edited volumes, monographs and book chapters in the Book Citation Index (BKCI) and Science Citation Index (SCI, SoSCI, A&HCI). *Journal of Scientometric Research*, 1(1), 28–34.

Moed, H. (2005). *Citation analysis in research evaluation*. New York: Springer.

Nederhof, A. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1), 81−100.

Testa, J. (2011). *The book selection process for the Book Citation Index in Web of Science*. Retrieved July 08, 2013, from http://wokinfo.com/media/pdf/BKCI-SelectionEssay_web.pdf

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E. y Delgado López-Cózar, E. (2012). Towards a 'Book Publishers Citation Reports'. First approach using the 'Book Citation Index'.

Revista Española de Documentación Científica, 35(4), 615-620. Retrieved July 18, 2013, from http://arxiv.org/ftp/arxiv/papers/1207/1207.7067.pdf

Torres-Salinas, D., Rodríguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J., & García, J. A. (2013). Mapping citation patterns of book chapters in the Book Citation Index. *Journal of Informetrics*, 7(2), 412-424.

Travis, H. (2010). Estimating the economic impact of mass digitization projects on copyright holders: Evidence from the Google Book search litigation. *Journal of the Copyright Society of the U.S.A.*, 57(4), 907-949.

Vincent, L. (2007). Google Book search: Document understanding on a massive scale. *Paper presented at the Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2819-823.

Weiss, A. & James, R. (2013). Assessing the coverage of Hawaiian and pacific books in the Google Books digitization project. *OCLC Systems and Services*, 29(1), 13-21.

White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, 60(6), 1083-1096.