Search Markets and Search Results: The Case of Bing[1]

Mike Thelwall, David Wilkinson
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.

Bing and Google customise their results to target people with different geographic locations and languages but, despite the importance of search engines for web users and webometric research, the extent and nature of these differences is unknown. This study compares the results of seventeen random queries submitted automatically to Bing for thirteen different English geographic search markets at monthly intervals. Search market choice alters a small majority of the top 10 results but less than a third of the complete sets of results. Variation in the top 10 results over a month was about the same as variation between search markets but variation over time was greater for the complete results sets. Most worryingly for users, there were almost no ubiquitous authoritative results: only one URL was always returned in the top 10 for all search markets and points in time, and Wikipedia was almost completely absent from the most common top 10 results. Most importantly for webometrics, results from at least three different search markets should be combined to give more reliable and comprehensive results, even for queries that return less than the maximum number of URLs.

Introduction

According to Alexa.com (www.alexa.com/topsites, Nov. 13, 2012), web search is central to 12 of the world's 25 most visited websites: Google (rank 1), Yahoo! (4, search and portal, now owned by Microsoft and with results driven by Bing), Baidu (5), Google India (14), Yahoo! Japan (15), Google Germany (17), Yandex (18), MSN (19, portal and Bing search), Google Hong Kong (21), Google Japan (22), Bing (23) and Google UK (25). Although the broad details of how search engines work are known, the details of their operations, and particularly the ranking of results and spam filtering, are unknown and seem to be closely guarded commercial secrets. Whilst it is known that the same query will generate different results over time and between search engines, the same query can also generate different results at the same time for different users based upon their geographic location or search preferences but the nature and extent of this variation is unclear.

Some search engines, including Bing, segment users into *search markets* when calculating the results of their query. These search markets are based upon geographic location and language. For example, one Bing search market is *English-India;* people in India searching Bing with English as their default setting will get different results to people searching in English in the US (the *English-USA* market). The search market is chosen by Bing, with users having the option to override it by, in November 2012, clicking the *Preferences* icon on the Bing.com home page and then the *Change your country/region* link. Bing's list of 40 options includes 10 that specify a language (the three for English were: Arab countries, Canada, and United States, all areas with other popular languages spoken). In contrast, Google seemed to have more fine-grained location based results (perhaps partly for its map vertical) since its results pages in November 2012 had a *Change location* link with

[1] Wilkinson, D., & Thelwall, M. (in press). Search markets and search results: The case of Bing. Library and Information Science Research.

a free-text field that recognised individual towns. Google also apparently has 188 different national or regional variants branded by domain name, such as Google.co.uk (http://www.google.com/supported_domains, as of November 23, 2012). For both search engines the regional results include international results and the user is able to separately request that *only* results from a certain country or region are returned. There seems to have been no research into the impact of search markets on search engine results, however.

Problem statement

Differences between search results are of particular interest in the field of webometrics, which often involves counting web search matches for large sets of queries. Many studies need lists of URLs matching a search that are as complete as possible, or hit count estimates (figures reported near the top of a search results page estimating the total number of results) as proxies for these (Ortega & Aguillo, 2009; Park, 2010; Spörrle & Tumasjan, 2011; Thelwall, Klitkou, Verbeek, Stuart, & Vincent, 2010; Vaughan & You, 2010). Until mid-2012, Bing hit count estimates could sometimes be used to avoid collecting complete URL lists, but they can no longer be accessed automatically (this is an undocumented side-effect of the early 2012 Bing API (Applications Programming Interface) 2.0 upgrade) and Google does not allow automatic searching. Hence all webometric applications requiring automatic Bing searches now rely upon URL lists matching a query, but this does not work if too many results are available for any given query because Bing truncates its results at 1000 matches. The query splitting technique (Thelwall, 2008a) can be used to gain extra matching URLs but this method is fragile due to its reliance upon selecting terms from search result titles and descriptions. If there are substantial differences between search markets then the total number of URLs matching a query could be increased instead by combining the results of multiple searches for different markets.

This article evaluates the extent to which the results from Bing English search markets differ for the first 10 results and for the complete set of URLs returned. Presumably search results for different languages are radically different but it is not clear whether this is true for results in the same language. Bing is used because, unlike the more popular Google, it seems to provide unfiltered access to search market results (described below in more detail), and because it can be used for automated queries in webometric studies.

The findings will inform those teaching search engine use and those relying upon search engines by revealing the extent of search engine geographic variations, at least for one major search engine. In particular, the findings will give preliminary insights into whether a search is likely to return the universally best matches (at least as determined by the search engine) for a query rather than a local interpretation of the best matches. To give a specific example, if a media student investigated critical reactions to an internationally distributed movie by searching for online reviews then they would need to know whether the reviews returned by the search engine were heavily influenced by the locality from which they searched. This would help them to interpret the results or to recognise the need for additional search steps to minimise any geographic bias in their results.

The findings will also improve the power of future webometric studies that rely upon Bing results by suggesting search market strategies to increase the amount of data (i.e., the number of matching URLs) that can be analysed. These strategies have now been embedded in the Webometric Analyst free software, making it more powerful.

Search engines and search results

Although the performance and algorithms used by the major commercial search engines are not public, some general information is known about how search engines work from publications (Brin & Page, 1998) and patents (Page, 2001) produced by their architects. In addition, some information science research has investigated the output of search engines, typically focussing on variations in results over time.

Search engine design

Search engines are complex distributed computing systems. In terms of architecture, a major search engine may have several different copies of itself, each being a collection of over 1,000 PCs that are connected together to efficiently search and return results to users (Badue et al., 2012). Queries submitted to a search engine are presumably directed to the local copy or to the copy that is currently the least busy. The local copies may be identical except for periods when the index is being updated. Search engines contain a general index in addition to verticals for aspects such as news, images, maps and videos. Each vertical searches a particular type of document and integrates the results into the main search, when relevant (Ka et al., 2010). For example, the news vertical has the task of integrating news into the main results, when relevant, and judging relevancy in a time-dependant way (Diaz, 2009). The news vertical is presumably updated frequently and so news-relevant searches may change from minute to minute for a fast moving story but the typical search may have no news-relevant matches.

Search engines find pages using web crawlers that start with a list of known web pages and then identify new pages iteratively by downloading known pages and extracting their hyperlinks to find new pages. Once a page has been found, crawling seems to be conducted irregularly and repeatedly to identify changes in pages (Lewandowski, 2008). Crawlers do not cover the whole web (Lawrence & Giles, 1999) because it is impossible to find all web sites (Thelwall, 2002) and perhaps also due to storage limitations. All search engines index different sets of web pages due to differences in crawling strategies, and differing amounts of historical data. Another important source of potential variation is in the algorithms used to identify and remove duplicate or near duplicate matches from search results (Thelwall, 2008b). Perhaps most importantly, however, search ranking algorithms seem to be the major source of differences between search engines in practice because users tend only to view the first page of results and so the method used by a search engine to select these is critical (Spink & Jansen, 2004). Historically, ranking has been based on the extent to which the query matches the pages (i.e., the traditional information retrieval approach) as well as using hyperlink counts to estimate the importance of pages (Brin & Page, 1998) but many other factors are probably used (Google, 2012) and search engines can also customise results to individuals based upon their search interests (Teevan, Dumais, & Horvitz, 2005) or of the interests of people like them (Weber & Castillo, 2010).

Finally, search engines are fundamentally commercial entities and design decisions can be driven by corporate needs rather than just by technical considerations, although the need to give users a good service is very important (E. Van Couvering, 2012; E. Van Couvering, 2007) .

Search engine results evaluation

There is a considerable amount of research that evaluates various aspects of search engines from an information science perspective and even an entire book on the topic (D. Lewandowski, 2012) but there seems to be no research that compares results across different search markets. This section focuses on analyses of the results returned by search engines rather than other aspects, such as hit count estimates (Uyar, 2009b) and changes in web pages (Koehler, 2004) or web content relevant to a topic (Bar-Ilan & Peritz, 2009), that are not directly relevant to this article.

Search engine results for the same query are known to change over time (Rousseau, 1999) and various methods have been developed to measure this (Bar-Ilan, 2002). Perhaps particularly worrying is that search engines sometimes do not return pages that are relevant to a search even though they have previously indexed them or are known to currently index them (Mettrop & Nieuwenhuysen, 2001). In some cases pages containing information that is not present in any other document returned can be omitted (Bar-Ilan & Peritz, 2008), resulting in a loss of information to the searcher. The first results page is probably most relevant to the typical user, however. An investigation of MSN (now Bing), Google and Yahoo! in 2006 found that it took 3 months for 50% of MSN result URLs to disappear from the list (for a popular query) but for Google and Yahoo! the time period was over a year (McCowan & Nelson, 2007), showing that very different updating strategies may be used.

Several studies have compared the overlap of results between different search engines. For example, a comparison between MSN Search, Google, Yahoo! and Ask Jeeves in 2005 using over 20,000 genuine user queries found that 84.9% of the URLs in the first results page were not found in the first results page of any of the other three search engines (A. Spink, Jansen, Blakely, & Koshman, 2006).

Although the major search engines attempt to maintain an international index, US sites have been found to be better covered than others (Pirkola, 2009; Vaughan & Zhang, 2007), perhaps because of the early entry into the Web of the US (Vaughan & Thelwall, 2004). This may affect the extent to which the results of different search markets vary if US results are common to many.

The most user-centred way to compare search engines is to use human judges to rate the quality of the results returned by each one for the same search, although this is time-consuming and expensive to do well. The human coder approach seems to be used by major search engines for their internal purposes. A published study from Yahoo! compared the first 5 results for Google, Microsoft Live Search, and Yahoo! in 2008 based upon 1,000 random Yahoo! queries (Zaragoza, Cambazoglu, & Baeza-Yates, 2010) and found that each search engine performed better than the others for a significant number of queries and so it makes sense for users to switch search engines for individual searches when getting poor results for a particular query.

One previous study has included a comparison of search markets for a search engine. This included a comparison of the top 10 results for 10 English language queries between Google.com and Google.co.il twice daily for 20 days in January 2004 (Bar-Ilan, Levene, Mat-Hassan, 2004). A very high average overlap in the top 10 results between the two versions of Google was found (apparently over 9.6 out of 10 common results in most cases) as well as a high rank correlation (apparently over 0.944 in all cases) except that the single query Web data mining gave much less consistent results (an average overlap of 8.3 and an overlap of only 3 for at least one time period). In this case, Google presumably detected the

submission language and returned results in the same language or the overlap would have been substantially different.

In partial summary, here are some reasons why two identical searches may return different results:

- Different search engine
    - Search engines index different parts of the web.
    - Search engines have different algorithms to rank the results and to filter out spam and near-duplicate results.
- Different users
    - Search engines may personalise results based upon the users' previous queries.
    - Search engines may categorise users into market segments by geographic location or interest types and simultaneously customise the results for whole groups.
- Different query submission times
    - News components may change rapidly.
    - The search engine algorithm may change or its parameters may change.
    - New relevant pages may be created or be found, and old pages may disappear or change to be irrelevant.
- No differences (i.e., the same query submitted by the user twice from the same computer and at the same time).
    - The queries may be directed to different copies of the search engine.
    - There may be a random parameter in the ranking algorithm that deliberately changes some results.

There are also many reasons why a page may not be returned in a results set for a given query, including the following.

- The search engine judges the page irrelevant (e.g., because it does not contain the query term(s)).
- The search engine has not found or has not indexed the page or the part of the index containing the page is temporarily broken.
- The page is a duplicate or local near duplicate of a previously-returned matching page
- The page is part of the same web site as at least two previously-returned matching pages.
- There are too many results to return them all to the user and the page is ranked too low to be delivered.

Research questions

The research questions concern the extent of variation of the top 10 results and all results returned for a query. The questions concerning the top 10 are most relevant to typical users that may not visit any more results and the questions concerning all URLs are most relevant to webometric studies, although in both cases the results may vary for different types of query.

- In terms of the overlaps between the results sets for the same query, do the top 10 and all search results vary more over time or more between search markets?
- What is the typical extent of overlap between search markets for the top 10 URLs and all URLs returned?
- Do search markets group together by cultural similarity or geographic proximity?

Methods

The overall research design was to conduct a series of identical searches in different search markets at a series of different points in time and to compare the results for the extent of overlap between them, using the Bing API 2.0 as the data source. The Bing API allows programmers to access the Bing search engine on a limited basis. The choice of the API was partly because it is used in webometric research and partly to ensure reliable results. An alternative way to collect the data would be through the normal Bing web interface but the use of this would risk the results being affected by Bing attempts to personalise the results to the user. It seems that the Bing API should give results that are less personalised and therefore reflect more the underlying logic of Bing markets. Search engine APIs in the past have given similar results to the web interface in terms of freshness but may be based upon smaller indexes (McCowan & Nelson, 2007). The free program Webometric Analyst was used to submit the queries to Bing and then to collect, process and compare the results for the different search markets.

A set of queries was needed for the comparisons. The results seem likely to depend upon the types of queries used. For example, academic queries should give more consistent results than commercial queries, since knowledge seems to be less geography-dependant than shopping. Hence, a random selection of queries was chosen as a baseline, accepting that the results would vary for other types of query. The set of queries chosen was a random collection of 20 moderate frequency words extracted using a random number generator from queries for Tweets in English in 2012, a convenience source of differing frequency words. The use of moderate frequency words from a web-based source seems likely to give search matches with enough results per search to allow reasonable comparisons of coverage. Common words (e.g., and, but) seem unlikely to be the subject of serious web searches. The choice of 20 keywords allowed all the searches to be run within the Bing rate restriction for free API use. For the 13 search markets these 20 terms will produce a maximum of 20*20*13=5,200 queries since each term can be submitted up to 20 times to return each page of 50 results. This figure is above the maximum monthly limit of 5,000 queries but this limit was not reached because most search terms returned less than 20 pages of results. Three queries were subsequently rejected because some of the search markets returned no results for sex-related searches. Hence, to prevent the comparison being affected by blanket bans in this way, three of the 20 queries were removed (cock, sickass, stripperella), leaving 17 for the comparison (blvd, doco, mcgowan, ipocalypse, pepto, marieke, havok, dotn, gurion, fitzwilliam, e-cover, tip-toe, thudangi, barocca, mudi, cahi, ball-point).

Although the queries were chosen to give appropriate numbers of results for the analysis they are not representative of the types of queries used in webometric research or of queries submitted by Bing users. Webometric queries tend to be complex, such as URL citation searches (Kousha & Thelwall, 2007), title mention searches, or keyword queries (Vaughan & You, 2010), and there are many different ways in which people search the web. Hence, it would not be possible to construct a set of query terms that would be in any way representative of the different uses. The terms chosen here will nevertheless give a baseline to indicate how search markets might perform but future research may reveal differences for specific types of search and even for different search topics (e.g., Thelwall, 2011).

The searches were conducted in English to allow the most comparisons because English is a common web language and Bing supports 13 different search markets for English in comparison to just 6 for Spanish, 4 for French, 3 for German and Chinese, 2 for Dutch and

Portuguese and 1 for every other supported language. The full set of 13 English search markets (as of August 2012) were chosen (using their Bing codes http://msdn.microsoft.com/en-us/library/dd251064.aspx): en-AU (Australia); en-CA (Canada); en-GB (UK); en-ID (Indonesia); en-IE (Ireland); en-IN (India); en-MY (Malaysia); en-NZ (New Zealand); en-PH (Philippines); en-SG (Singapore); en-US (USA); en-XA (Arabia); en-ZA (South Africa).

The searches were submitted three times, at monthly intervals. Monthly intervals were chosen as in the past search engines seem to have rolled out major search index updates every month and so an interval of a month should capture such changes. Bing does not officially report details of its search operation and so this is speculation.

Step by step methods instructions for data collection and initial processing are available here: http://lexiurl.wlv.ac.uk/searchmarkets.html.

Results

The results of the analysis of the seventeen queries is organised below by research question.

The first question asked whether search results varied more over time or more between markets. The top 10 results between *different* markets at the same point in time overlap (diagonal values in Table 1) approximately the same amount as between the *same* market at different points in time (off-diagonal values in Table 1), at least for gaps of one or two months. Although the overall difference is significant at $p<0.001$ using an independent samples t-test, the overall Jaccard similarity difference is only 0.088. The exception is that between the second and third month, the similarity within markets at the different months was much higher that the similarity between markets at the same point in time. In contrast, for all results the similarity between different markets at the same point in time (diagonal values in Table 2) is higher than the similarity for the same market at different points in time (off-diagonal values in Table 2). This is clearest from the Jaccard similarity scores since the total number of results varied between months. The overall Jaccard similarity difference of 0.235 is substantial and significant at $p<0.001$ using an independent samples t-test. The statistical tests reported here are indicative rather than robust because of possible changes in Bing between months.

Table 1. Average overlap (top) and Jaccard similarity (bottom) between the *first 10 results*. Diagonal values indicate the average overlap/similarity between different markets at the same point in time (n=13x13-13), whereas off-diagonal elements are the average overlap/similarity for the *same* market at different points in time (n=13).

Data set (av. URLs)	1st	2nd	3rd
1st (10)	4.135 0.280	4.797 0.318	4.611 0.301
2nd (10)		4.714 0.339	7.200 0.571
3rd (10)			4.796 0.346

Table 2. Average overlap (top) and Jaccard similarity (bottom) between *all results*. Diagonal values indicate the average overlap/similarity between different markets at the same point in time (n=13x13-13), whereas off-diagonal elements are the average overlap/similarity for the *same* market at different points in time (n=13).

Data set (av. URLs)	1st	2nd	3rd
1st (625.8)	429.0 0.455	244.3 0.274	279.4 0.279
2nd (509.1)		352.1 0.459	316.8 0.372
3rd (655.0)			495.8 0.534

The second research question addressed the typical extent of overlap between search markets. The average overlaps in the top 10 URLs between markets for each month are 41.4%, 47.1% and 48.0% (the diagonal values in Table 1 to one decimal place). The overlap between two sets of query results is defined to be the number of URLs that are common to both and the average overlap between multiple sets of results is the average of the overlaps. This suggests that, on average, just under half of the top 10 results are the same between different markets. For all URLs the overlaps between markets for each month are 68.6%, 69.2% and 75.7% (dividing the diagonal figures in Table 2 by the average number of total results for each month). Hence, over two thirds of all matches returned are the same (Table 2) and the overlap is consistently larger than for the top 10 results.

In terms of the overall overlap between results for the same query, Figure 1 illustrates the typical composition of the top 10 results for a query. This shows that rare URLs are more frequent than common URLs. For instance, 4.5% of the URLs for a given query did not occur in the top 10 for any other search market or even for the same search market in one of the other two months (i.e., 1 on the horizontal axis of Figure 1.) In contrast, 0.5% of the URLs in the top 10 occurred in all 39 results for the same query. The situation is different for the complete sets of results (Figure 2): over 5.5% of the results were returned for all 39 queries but only 1.5% of URLs in any given complete set of results do not appear in the results of any of the other 38 queries for the search term.

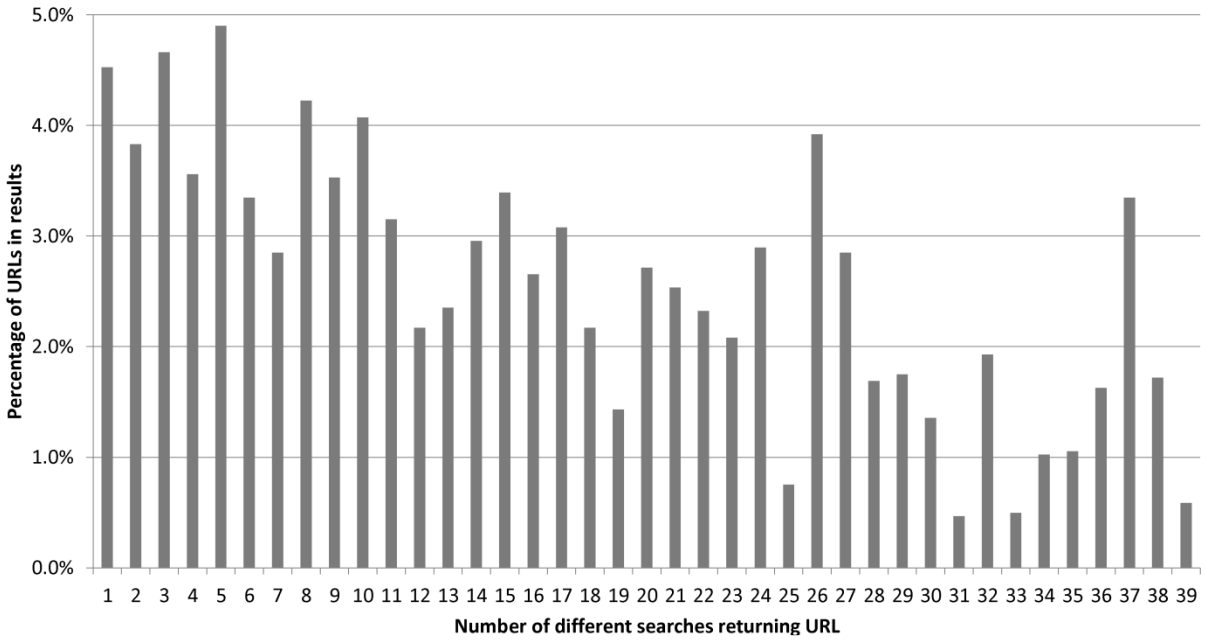

Figure 1. The percentage of URLs of each frequency in the typical top 10 results for a query. For example, an URL with frequency 25 occurred in the top 10 for 25 of the 39 searches for a query; from the graph, about 0.8% of the URLs in any given top 10 results set occurred in 25 of the top 10 results sets.

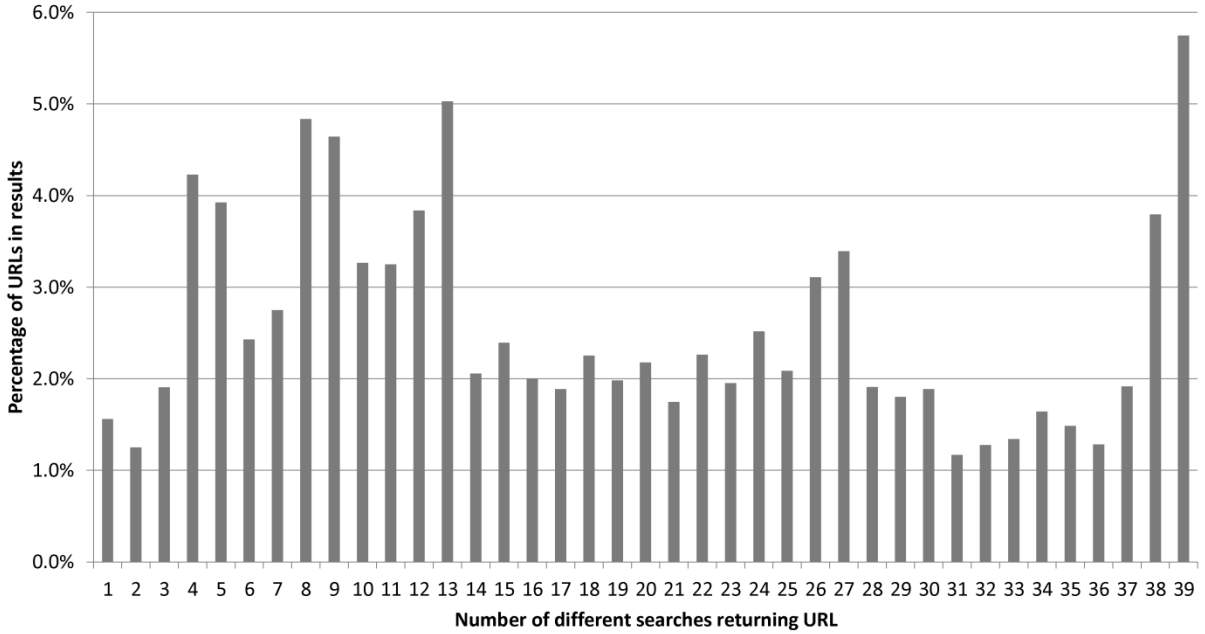

Figure 2. The percentage of URLs of each frequency in the typical complete set of results for a query. For example, an URL with frequency 25 occurred somewhere in the results for 25 of the 39 searches for a query.

The third research question asked whether markets grouped together by cultural similarity or geographic proximity. Figures 3 and 4 illustrate the degree of commonality between English markets for the 17 queries using Multi-Dimensional Scaling (MDS) based upon Jaccard dissimilarity scores (a measure of the extent of overlap between the results sets) for each of the three time periods separately, then averaged over the 3 time periods.

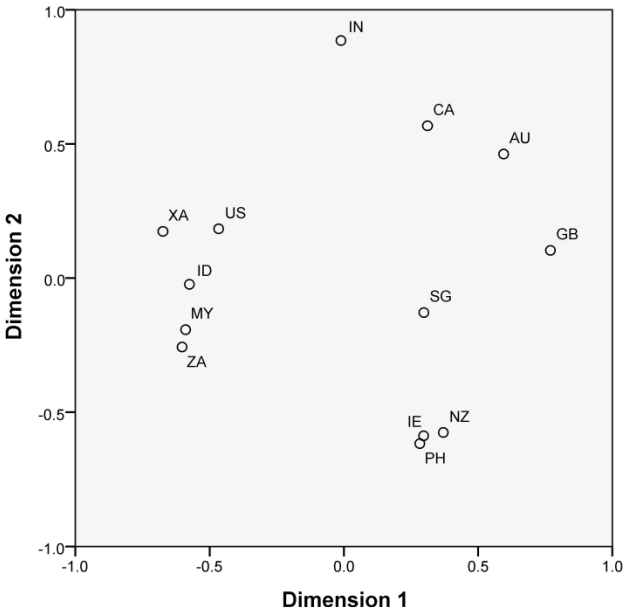

Figure 3. MDS diagram (PROXSCAL) of commonality between the *top 10* results of the 13 English search markets, using Jaccard dissimilarity, and averaged over the 3 data collection periods. Normalised stress: 0.047.

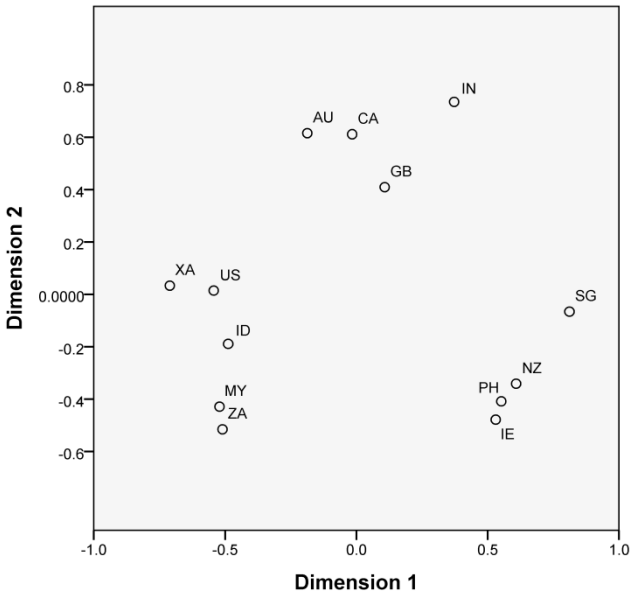

Figure 4. MDS diagram (PROXSCAL) of commonality between *all* results of the 13 English search markets, using Jaccard dissimilarity, and averaged over the 3 data collection periods. Normalised stress: 0.028.

The results suggest, somewhat counter intuitively, that there are three main clusters of search markets that do not seem to be based upon geographic proximity or overall cultural similarity: Group 1 is XA, US, ID, MY, ZA; Group 2 is AU, CA, GB, IN; Group 3 is IE, PH, NZ and possibly SG, although SG could also be an isolate.

Discussion

There is a surprising amount of variation between the URLs returned for the different English search markets, as returned by the Bing API. This finding should not be taken as

indicative of all queries because of potential variations due to different types of search, such as academic, educational, cultural or commercial queries. Indeed, it seems likely that academic queries would exhibit less variation and cultural and commercial queries would exhibit more variation. Similarly, the results may be different for popular queries and for complex queries, such as multi-word or phrase searches. Another limitation is that the extent of overlap between search engines may be larger than it appears to be if similar pages are returned with different URLs for different search markets.

One possible source of variation between the results is in the inclusion of news vertical results. To check for evidence of this, we searched the URLs for two major English language news portals: CNN and BBC, finding no contemporary BBC news URLs in any top 10 (although one old news page from 1998 was included in 11 results sets), and no contemporary CNN news URLs in any top 10 (although one page from 2008 and one from 2011 was included). In the full set of URLs, there were some current news stories. One BBC page published on November 1, 2012 occurred in 8 of the 13 full sets of results for the final round of searching, but only 36 of the 40,841 URLs were from the BBC and only 23 were from CNN, suggesting that news was not a major feature of any of the results sets, even allowing for the inclusion of a small number of pages from a range of other English language news sources.

For the top 10 results, only one URL was universal in the sense of being found in every top 10 for all search markets and all three time periods. This URL was a frequently asked questions page for the brand name product Pepto Bismol, maintained by its manufacturer:
http://www.pepto-bismol.com/pepto-bismol-faq.php (for the query *pepto*). Three URLs were present in 38 of the 39 results pages (query *havok*: http://havokband.com/
and http://www.comicvine.com/havok/29-3546/; query *fitzwilliam*: http://www.fitzwilliamhotel.com/). Given that all the queries were in English, it seems surprising that English Wikipedia was not common to many queries. Only one Wikipedia page was almost ubiquitous: http://en.wikipedia.org/wiki/Havok_(comics) for havok (37 results, absent from the ZA and MY top 10 results in the first month). The next most common Wikipedia page was http://en.wikipedia.org/wiki/Ballpoint_pen for ball-point (28 results, never returned for the top ten for AU, CA or IN; missing from the first month for US and SG).

To confirm that national factors were important in the URLs returned for the top 10 results we checked each URL that was returned twice or three times for one search market top 10 but never in any other search market. Of the 87 URLs matching this condition, a majority (52) originated in the region associated with the search market returning them. The remainder were mainly US sites or sites based in the US with an international focus (e.g., dictionary.com) although there were a few odd cases, such as a South African site only occurring in the Indian search market results.

For the complete sets of results, 583 URLs were ubiquitous, occurring in all sets. Surprisingly, none of these were Wikipedia pages, although one was from another Wikimedia foundation site (http://en.wiktionary.org/wiki/Pepto-Bismol for *pepto*). Some of these terms did not have a natural single relevant Wikipedia page, which is a partial explanation of these results. Also perhaps surprisingly, one page was from an anti-semitic site (http://www.jewwatch.com/jew-leaders-ben-gurion.html for *gurion*) and one was from a pornographic site containing a disclaimer about the adult contents to be found inside

(http://stonecatfights.com/celebrity/barocca.html for *barocca*). Six of the 583 pages were from US universities.

To set the results in context, the 17 queries were rerun three times by three different people for the en-GB location within the space of 17 minutes. These three near-simultaneous API results were also compared to the first page of results for the same 17 queries submitted to Bing through a web browser from the UK, accessed during the same hour. For two of these queries, Bing offered the option to redo the search after disabling the term auto-correction (e.g., changing dotn to dont) and this option was taken as it gave similar results to the API.

The three near-simultaneous API top 10 results were similar but not identical, with an average overlap of 9.02 out of 10 and some ranking changes (Table 3, column 3). The top 10 had an identical ranking for less than half (7) of the queries and in one case even the first result was not always the same for a query. When the results differed, the cause was usually an URL occurring in one or two of the results that did not occur in the other(s). The web results were compared to the third API results (Table 3, column 2) and in no cases were the top 10s identical; the average overlap was 7.4 between the web search and the third API results. In three cases the only difference between the Web and API results was the order (cahi, blvd, barocca) but in the remaining 14 cases the URLs were different, suggesting that the Web ranking is closely related to the API ranking but introduces some new factors and beyond the top few results the list changes considerably. Two obvious differences were that the API contained only one news vertical result in the query that had these, whereas the Web browser query had three news vertical results; and two web queries contained image vertical results but the API results contained none.

For the complete set of queries, all three queries returned identical complete sets of results in only one case (Table 3, column 4). The average Jaccard similarity between pairs of results was 0.159, much lower than the equivalent figures in Table 2. Despite the short time period, there seem to be changes over time in the results sets rather than random changes (e.g., due to alternating between multiple non-identical copies of Bing) because the most similar rankings were either the first two or the last two results sets but never the first and the last. The most common cause of differences between results was the introduction of a new URL in one set that was not in the other set rather than a ranking change. This suggests that URLs were occasionally added to the results and sometimes removed. Some of the URLs were new (one was a day old) but others were old enough to seem likely to have been previously known to Bing, with one being apparently four years old. In 7 cases the two most similar results sets were identical. Most worryingly for webometrics, in one case the result set of one API query stopped prematurely (ipocalypse), giving the impression that this query had few results but the other identical queries produced many more results. Also, in two cases (doco, dotn) two results sets stopped prematurely in an identical fashion. A clear implication for future webometrics research is that the total number of URLs returned for a query is unreliable because of occasional premature truncation; for reliable results, a query should be repeated at least 3 times, taking the largest figure or combining the results.

Table 3. A comparison of key differences between the results of 3 near simultaneous API searches and a web browser search for 17 queries.

Term	Web ranking vs. final API ranking	Initial number of identical results for queries 1, 2 & 3; source of first difference	Biggest initial number of identical results for two queries; source of first difference
cahi	5&6 swapped	485 (all results)	485 (all results)
blvd	7&8, 9&10 swapped	35; URL in 3 not in 1&2	49 (1&2); ranking order change
fitzwilliam	After 2 mildly different	16; URL in 3 not in 1&2	424 (1&2); URL in 2, also in 3
mcgowan	1&2, 4&6,7&9 swapped; 10 new	13; URL in 2&3 not in 1	651 (2&3, all results)
gurion	3&4&5 swapped, after 8 very different	12; URL in 3 not in 1&2	13 (1&2); URL in 1 not in 2&3
dotn	After 9 very different	12; URL in 2&3 not in 1	48 (2&3, all results; 1 has 461 results)
ball-point	After 2 very different	11; URL in 1 not in 2&3	740 (2&3, all results)
doco	1&2, 7&8 swapped, 10 new	9; ranking position swap	49 (2&3, all results; 1 has 572 results)
havok	After 7 very different	7; URL in 2&3 not in 1	186 (2&3); URL in 1 not in 2&3
marieke	1&2 swapped, after 6 very different	6; URL in 3 not in 1&2	12 (1&2); URL in 1 not in 2&3
mudi	After 5 very different	6; URL in 2&3 not in 1	486 (2&3); URL in 2, also in 1
pepto	After 6 very different	6; URL in 2&3 not in 1	287 (2&3); URL in 3 not in 1&2
e-cover	1&2 swapped, after 2 very different	5; URL in 2&3 not in 1	839 (2&3, all results)
ipocalypse	After 2 very different	2; different URLs from same site	48 (2&3); 2 stops, 3 continues for 273 more
tip-toe	After 2 very different	2; different URLs for same page	493 (2&3); URL in 3, also in 1
thudangi	After 1 very different	1; URL in 2&3 not in 1	12 (2&3); URL in 3, also in 1
barocca	7&8&9 swapped	0; ranking swap for first 2	909 (2&3, all results)

Conclusions

The top 10 results for the queries tested here showed substantial variations with the average overlap between any pair of search markets being less than 50%. There was more overlap between the full sets of results, with a majority of URLs being the same, on average, between pairs of different search markets. The extent of overlap between different search markets' results was about the same as the overlap for the same market a month later for the top 10 results, but less for the complete set of results. As expected, the search markets clustered into groups in terms of the extent of overlap for the URLs returned but the clusters did not conform to geographic proximity or cultural similarity. Web results seem to be initially similar to the API results but sometimes have substantial differences after a point in the top 10. They seem to be more dissimilar to the API results than the API results are to each other, suggesting that the web interface has some fundamental differences from the API.

From a webometric perspective rather than a search overlap perspective, the most important factor is the total number of URLs found matching a search. For the seventeen queries used, the average number of matches for a single query was 597 but the average number of URLs matching each query by combining the different markets and the three

different points in time is four times larger at 2,381. For the three time periods, combining the results of all 13 search markets would at least double the results from 626 to 1,473 from 509 to 1,223 and from 655 to 1,211 for the three different months. This confirms that running a query over multiple search markets and at different points in time will substantially increase the total number of URLs found matching a search, as required by some webometric research. In addition, since API results sometimes truncate without warning, future research should use at least 3 identical results to minimise the risk of a set of results being contaminated by truncation. This recommendation has now been implemented in the Webometric Analyst software (http://lexiurl.wlv.ac.uk/searchMarketsMultipleQueries.html).

Since search engine parameters change without notice or warning, the numerical results here are unlikely to be valid for a long period of time but the results can serve as a baseline and the method described in this article can be used to reassess changes and to compare the results for different types of search. For instance, it would be useful to know how typical student search results varied between markets. It would also be valuable to compare search results for different national Google variants using its normal web interface. Preliminary testing suggests that Google varies more than Bing by country for some searches (e.g., Marieke) but less for others (e.g., ball-point) and Wikipedia is more prominent in its results. Google seemed to incorporate UK results (the search location) even when using non-UK Google variants (e.g., www.ballpoint.co.uk appeared in response to all the ball-point queries tested) and so a proper experiment may need to recruit testers in different countries.

Acknowledgement

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. The funding source had no role in the study, including: design, the collection, analysis and interpretation of data; the writing of the report; and the decision to submit the article for publication.

References

Badue, C., Almeida, J., Almeida, V., Baeza-Yates, R., Ribeiro-Neto, B., Ziviani, A., & Ziviani, N. (2012). Capacity planning for vertical search engines. *ArXiv*, Retrieved July 22, 2010 from: http://arxiv.org/abs/1006.5059.

Bar-Ilan, J. (2002). Methods for measuring search engine performance over time. *Journal of the American Society for Information Science and Technology, 53*(4), 308-319.

Bar-Ilan, J., Levene, M., Mat-Hassan, M. (2004). Dynamics of search engine rankings – A case study. In: Proceedings of the Third International Workshop on Web Dynamics, ACM Press, New York, NY (pp. 3-13).

Bar-Ilan, J., & Peritz, B. C. (2008). The lifespan of 'informetrics' on the web: An eight year study (1998-2006). *Scientometrics, 79*(1), 7-25.

Bar-Ilan, J., & Peritz, B. C. (2009). A method for measuring the evolution of a topic on the web: The case of “Informetrics”. *Journal of the American Society for Information Science and Technology, 60*(9), 1730-1740.

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.

Diaz, F. (2009). Integration of news content into web results. *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM09),* Barcelona, Spain. 182-191.

Google. (2012). Google basics: Learn how google discovers, crawls, and serves web pages. Retrieved Nov 13, 2012, from http://support.google.com/webmasters/bin/answer.py?hl=en&answer=70897

Ka, C., Cuncannan, H., Wong, K., Steele, M., Gordon, M., Prakash, S., & Bergstraesser, T. (2010). In Microsoft Corporation (Ed.), *Consumer-focused results ordering* (2007/0038620 A1 ed.). US:

Koehler, W. (2004). A longitudinal study of web pages continued: A report after six years. *Information Research, 9*(2), Retrieved September 20, 2007 from: http://informationr.net/ir/9-2/paper174.html.

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

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature, 400*(6740), 107-109.

Lewandowski, D. (Ed.). (2012). *Web search engine research*. Bradford: Emerald.

Lewandowski, D. (2008). A three-year study on the freshness of web search engine databases. *Journal of Information Science, 34*(6), 817-831.

McCowan, F., & Nelson, M. L. (2007). Agreeing to disagree: Search engines and their public interfaces. *Joint Conference on Digital Libraries,* Retrieved June 18, 2007 from: http://www.cs.odu.edu/~fmccown/pubs/se-apis-jcdl07.pdf.

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation, 57*(5), 623-651.

Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management, 45*(2), 272-279.

Page, L. (2001). In United States Patent (Ed.), *Method for node ranking in a linked database*

Park, H. W. (2010). Mapping the e-science landscape in south korea using the webometrics method. *Journal of Computer-Mediated Communication, 15*(2), 211-229.

Pirkola, A. (2009). The effectiveness of web search engines to index new sites from different countries. *Information Research, 14*(2), Retrieved 20 May 2009 from: http://InformationR.net/ir/14-2/paper396.html.

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics, 2/3*, Retrieved July 25, 2006 from: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html.

Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the web*. Dordrecht: Kluwer Academic Publishers.

Spink, A., Jansen, B. J., Blakely, C., & Koshman, S. (2006). A study of results overlap and uniqueness among major web search engines. *Information Processing & Management, 42*(5), 1379-1391.

Spörrle, M., & Tumasjan, A. (2011). Using search engine count estimates as indicators of academic impact: Web-based replication of haggbloom et al.'s (2002) study. *The Open Psychology Journal, 4*(1), 12-18.

Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *Proceedings of the 28th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval (SIGIR 05), Salvador, Brazil. 449-456.
Thelwall, M. (2002). Methodologies for crawler based web surveys. *Internet Research: Electronic Networking and Applications, 12*(2), 124-138.
Thelwall, M. (2008a). Extracting accurate and complete results from search engines: Case study windows live. *Journal of the American Society for Information Science and Technology, 59*(1), 38-50.
Thelwall, M. (2008b). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology, 59*(11), 1702-1710.
Thelwall, M. (2011). A comparison of link and URL citation counting. ASLIB Proceedings, 63(4), 419-425.
Thelwall, M., Klitkou, A., Verbeek, A., Stuart, D., & Vincent, C. (2010). Policy-relevant webometrics for individual scientific fields. *Journal of the American Society for Information Science and Technology, 61*(7), 1464-1475.
Uyar, A. (2009b). Investigation of the accuracy of search engine hit counts. *Journal of Information Science, 35*(4), 469-480.
Van Couvering, E. (2012). Search engines in practice: Structure and culture in technical development. In G. Bolin (Ed.), *Cultural technologies: The shaping of culture in media and society* (pp. 118-132). London: Routledge.
Van Couvering, E. (2007). Is relevance relevant? market, science, and war: Discourses of search engine quality. *Journal of Computer-Mediated Communication, 12*(3), Retrieved May 14, 07 from: http://jcmc.indiana.edu/vol12/issue3/vancouvering.html.
Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management, 40*(4), 693-707.
Vaughan, L., & You, J. (2010). Word co-occurrences on webpages as a measure of the relatedness of organizations: A new webometrics concept. *Journal of Informetrics, 4*(4), 483-491.
Vaughan, L., & Zhang, Y. (2007). Equal representation by search engines? A comparison of websites across countries and domains. *Journal of Computer-Mediated Communication, 12(3), Retrieved May 14, 2007 from: http://jcmc.indiana.edu/vol12/issue3/vaughan.html*
Weber, I., & Castillo, C. (2010). The demographics of web search. *Proceedings of SIGIR 2010* (pp. 523-530). New York: ACM Press.
Zaragoza, H., Cambazoglu, B. B., & Baeza-Yates, R. (2010). Web search solved? all result rankings the same? *Proceedings of the 19th ACM International Conference on Information and Knowledge Management,* Retrieved February 3, 2011 from: http://www.hugo-zaragoza.net/academic/pdf/zaragoza_CIKM2010.pdf.