1-1-2004

Search engine coverage bias: Evidence and possible causes

L. Vaughan

Vaughan, L.; Fac. of Info. and Media Studies, University of Western Ontario, London, Ont. N6A 5B7, Canada; email: lvaughan@uwo.ca

M. Thelwall

Search Engine Coverage Bias:
Evidence and Possible Causes

Liwen Vaughan[*]
Faculty of Information and Media Studies
University of Western Ontario
London, Ontario, N6A 5B7, Canada
Phone: (519) 661-2111 ext. 88499
Fax: (519) 661-3506
E-mail: lvaughan@uwo.ca

Mike Thelwall
School of Computing and Information Technology, University of Wolverhampton,
35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK.
Phone: + 44 1902 321470
Fax: + 44 1902 321478
E-mail: m.thelwall@wlv.ac.uk

Abstract

Commercial search engines are now playing an increasingly important role in Web information dissemination and access. Of particular interest to business and national governments is whether the big engines have coverage biased towards the U.S. or other countries. In our study we tested for national biases in three major search engines and found significant differences in their coverage of commercial Web sites. The U.S. sites were much better covered than the others in the study: sites from China, Taiwan and Singapore. We then examined the possible technical causes of the differences and found that the language of a site does not affect its coverage by search engines. However, the visibility of a site, measured by the number of links to it, affects its chance to be covered by search engines. We conclude that the coverage bias does exist but this is due not to deliberate choices of the search engines but occurs as a natural result of cumulative advantage effects of U.S. sites on the Web. Nevertheless, the bias remains a cause for international concern.

Keywords: Search engine coverage bias, linguistic factor, cross-country comparison

Introduction

With the Web threatening to become an ubiquitous channel for delivering information, the interfaces through which users discover that information have a major economic and political significance. Commercial search engines are a logical and common first port of call for the discovery of Web resources. The temporary blocking of Google in China (BBC, 2002) is a case

---

where the political significance of search engine coverage has been recognised. However, we are more interested in looking at the issue from the opposite perspective. The question that we would like to ask is whether the coverage of the major search engines is biased along national or linguistic lines, and if so, what are the underlying causes of this bias. For the purpose of this study, "bias" is conceptually defined as a non-proportional coverage of a particular group of Web sites, e.g. sites from a particular country. Under representation of a group of sites by a search engine is considered a bias against these sites by the engine while over representation is viewed as a bias in favour of the sites in question. "Bias" is operationally defined (measured) as the percent of a group of sites that are covered by a search engine. Specifically, the coverage of a Web site was determined by comparing the number of Web pages indexed by the search engine with the total number of pages found by an independent crawl of the Web site in question. The study examined the coverage by Google, AltaVista and AllTheWeb of a sample of commercial Web sites from China, Singapore, Taiwan and the U.S. in order to test whether there are significant national or linguistic biases in their coverage.

Clearly, given the international nature of the Web, it would be a cause for concern for businesses if a user searching for an online product were only pointed to a given set of U.S. firms because the search engine that they used had not indexed other sites selling the same product. Such imbalances could be the cause of international political friction if the volume of online trading continues to grow. Fortunately, international Web users do not rely exclusively on the major U.S. based search engines. Many countries also have a range of home-grown alternatives (e.g. voila.fr for France, dino-online.de for Germany, and china.asiadragons.com for China). Nevertheless the international reach of the commercial giants is demonstrated by the proliferation of national and linguistic variants of these search engines. The variations can be in the interface alone, however, with the databases underneath being common to all versions (Sullivan, 2001). Logically, then, if a major search engine indexes a country's sites poorly but, say, only 10% of nationals use the engine then that 10% will have their online business directed internationally, but there will be little compensation through international traffic directed into the country by the same engine. The net result will therefore be a drain on the national economy. Another reason for the importance of search engine politics from a more sociological perspective is the consequent potential to "narrow the Web's functioning in society" (Introna & Nissenbaum, 2000) by marginalizing certain types of information, for example minority interests or pages in developing countries.

If search engines do display national biases then it would be useful to identify the underlying causes so that it could be predicted whether the bias would be likely to continue, whether search engine owners are deliberately fostering national imbalances or whether cultural differences were the cause and remedial steps could be identified to close the gap. Of particular concern is the indexing of Web sites that use non-ASCII character sets in case this is one of the factors militating against effective indexing. Grefenstette and Nioche (2000) showed that non-English language sites were growing more rapidly than English sites and so logically they could eventually dominate the Web, an issue that search engine designers must surely have had to consider. For this reason, major efforts are underway to tackle the problems of cross-language information retrieval (e.g. Peters, 2001; Oyama et al, 2003). Numerous studies have been carried out to deal with the difficulties of indexing and retrieval in Chinese (e.g. Nie & Ren, 1999).

Literature Review

There is a body of literature that discusses search engine retrieval performance from a purely information retrieval perspective (Hawking *et al*., 2001; Bar-Ilan, 2002). For example Gordon and Pathak (1999) evaluate search engine results through comparisons with human evaluations. One important aspect of a search engine quality is its coverage. Although Hawking et al (2001, p. 51-52) found no significant correlation between coverage and a specific type of precision they calculated, they did acknowledge that coverage could be very important or crucial for certain types of queries. The underlying assumption of our study is that the coverage is important and that an unbiased coverage is desirable. Mowshowitz and Kawaguchi (2002) studied search engine bias by testing whether one search engine retrieves results that are significantly different from those of a control group of search engines. This way of assessing the bias of an individual search engine would only be appropriate for our purposes if we could assume that the coverage of the control group of search engines was unbiased. One previous study compared search engine coverage of 42 different countries (Thelwall, 2000) in June-July 1999, concluding that great variations among countries were evident for all the engines tested: Yahoo!, Hotbot, AltaVista, MSN, and InfoSeek. For example, AltaVista covered 82% of the Finnish sample sites but only 36% of the Egyptian. The variations were reasonably consistent among search engines. However, no attempt was made in that study to statistically analyse the results in order to explain the discrepancies, although differing patterns of site construction were found and suggested as possible contributory factors. The study shows that the Mowshowitz and Kawaguchi (2002) technique is likely to be ineffective for evaluating national biases in search engines and so a different approach must be used in the current study.

Lawrence and Giles (1999) conducted a major investigation into search engine coverage of Web sites. Although they did not investigate national dimensions they showed that the major engines of the time covered no more than 16% of their sample and also that sites with more links directed at them (inlinks) were more likely to be indexed, an important finding. This could be intuitively expected, as they point out, since search engines can identify new sites to crawl by following links from previously crawled sites. They also found a significant overlap between search engines, but also large areas of non-overlapping coverage. The Lawrence and Giles (1999) results also rule out a potential technique for evaluating search engine biases: random walks (Henzinger *et al*., 1999) because of its reliance upon links.

As a side note on another aspect of search engine coverage, Huberman and Adamic (1999) claimed that search engines in 1999 indexed no more than 100,000 pages per site. National biases in total indexed sizes for sites would thus be hard to prove because it would be difficult to show that a search engine had a higher site limit for one country than another because it would involve finding and crawling a sample of very large sites. One paper has surveyed techniques for collecting data through crawlers, including random selection techniques (Thelwall, 2002a). It recommended random sampling of a subset of domains by domain name that should generate a relatively random samples of Websites that are not dependent on commercial search engines. This technique was adopted in this study and is described in detail below.

Methodology

The overall approach to the study is as follows. Define the countries and search engines to be compared; take a random sample of commercial Web sites from each chosen country; determine the size of each Web site (measured by the number of pages on the site) by a customized Web crawler; search for the Web site using each of the chosen search engine to find

out the number of pages of the Web site that is covered by the search engine; compare the data collected by the crawler with those collected from the search engine to determine if there is a significant difference in coverage by the search engines.

Since significant differences in coverage were found (details below), possible causes of the bias in the coverage were further investigated. Political and social causes would be difficult to measure and controversial. It would need a much larger scale and more complex and extensive investigation. Therefore, this study limited the examination of causes to technical ones. One possible cause is the technical difficulty of processing non-ASCII languages. Another possible cause is the visibility of a Web site. If the search engine does not know the existence of a particular site, it will not index it. Therefore, the most possible technical cause of the biased coverage is the visibility of a site. For the purpose of this study, the visibility of a site is measured by the number of links to the site from other sites. This is based on the fact that major search engines all use crawlers or spiders to build up their databases of Web sites. A crawler finds Web sites by following links on the sites visited, so the more links a site receives, the more visible the site is to the search engine and the more likely it will be indexed.

Countries in the Study

Four countries[1] were selected for the study: U.S., China, Singapore, and Taiwan. The selection of the U.S. was obvious because U.S. sites dominate the Web in that there are more U.S. sites than sites from any other country (OCLC Web Characterization Project, 2002). All major search engines originated from the U.S. so they are more likely to favour U.S. sites if there is such as a bias. China was chosen as a contrast to the U.S. The U.S. is a developed country while China is a developing country. Use of the Web for business purposes started later in China but is developing very fast. The technical requirements to process the Chinese language could be a factor that deters search engines from covering Chinese sites. To examine if language is a factor, Taiwan and Singapore were also chosen. They are both developing countries and culturally similar to China. However, English is the official language of Singapore while Chinese is the official language of Taiwan. To control the language variable, sites from China and Taiwan that were not in the Chinese language and sites from Singapore that were not in English were excluded from the study.

Sampling Methods

Random samples of commercial Web sites for the four countries were taken through a two-step procedure. Step 1 was to use a computer program to generate domain names randomly and then send a Web crawler to visit the site to record site size information. Step 2 was to manually filter all sites crawled to make sure that they fit the criteria of the study.

STEP 1: Sampling sites and crawling valid sites

In order to select a genuinely random sample of the commercial Web sites from a country, we would need a complete list of all sites from that country so that selections could be made using a random number generator. Unfortunately, given the nature of the Web no such list

---

[1] Taiwan is not recognized as an independent country by the United Nations and other international organizations. It is therefore more accurate to use the wording country/region here. However, for the convenience of reading and writing, we used the wording "country" throughout the paper instead of "country/region".

exists so modifications had to be made in order to get a pseudo-random sample. The main modification is that we will select only sites with an official national domain name. For China this is com.cn, for Taiwan com.tw, for Singapore, com.sg and for the U.S. com. This excludes sites with a shared domain name, such as those hosted by geocities.com. For example, the home page of one U.S. business is http://www.geocities.com/sbs_77049/sbshomepage.html and it shares the www.geocities.com domain name with probably millions of other businesses, organisations and individuals. Given the relatively low cost of domain names now, we would expect that most serious U.S. businesses would have bought their own domain name. Therefore the omission of shared domain names is not expected to affect the sample significantly. Efforts were made to make the sample as inclusive as possible. For example, the .com sites that were not U.S. were redistributed to the appropriate countries in the study (details below).

It is also not possible to obtain a complete list of domain names with a given top level domain name. Although this is technically feasible over the Internet (Albitz & Liu, 2001) the feature is typically disabled in the software that has the capability of giving the required information. Moreover, since lists of domain names would be of value to spammers, administering organisations tend not to give these out, although it has been done in the past (Ju-Pak, 1999). An alternative approach is random IP address sampling, (e.g. O'Neill et al., 1997; Lawrence & Giles, 1999) within the range allocated to a country, but the virtual server facility now makes this approach ineffective (Thelwall, 2002a). The virtual server HTTP facility was introduced partly to combat the problem of an insufficient number of IP addresses to allocate one to each domain name. Its widespread implementation now means that one IP address can correspond to any number of domain names. In particular, Web hosting companies can use just one IP address but host an unlimited number of domain names. This practice is so widespread that IP address sampling is no longer effective and IPv6, the replacement for the current IPv4 address system with a hugely increased number of IP addresses, has not been considered necessary (http://www.ipv6.org/). The approach that we adopted was to randomly generate legal domain names from a limited Web space and then test the domain names sampled to see if they were in use. To achieve a realistic success rate from this random guessing, these domain names have to be short. We limited our sampling frame to domain names with up to four letters based on the experience of a previous study (Thelwall, 2000). For example, for commercial sites from China, domain names with all the possible combinations of four letters, from www.a.com.cn to www.zzzz.com.cn, were included in the sampling. The computer program generated a random domain name in this range and then tried to retrieve its home page. If no response was received, then the domain name was discarded and the next randomly generated domain name tried.

The short domain name restriction could be a source of bias. Probably, short domain names tended to be used earlier than longer ones, so their Web sites would tend to be older. However, this would be true for every country so the comparisons among countries made in this study would not be seriously affected, except that this would be less of a problem for smaller countries such as Singapore. There is likely also to be some element of cultural and linguistic variation in this too: perhaps Web sites using a non-ASCII language would tend not to want a long domain name which could be difficult for users to remember. Nevertheless, given that it is not possible to obtain a genuinely random sample of Web sites (Thelwall, 2002a), we believe this to be an acceptable compromise.

Each valid site found was crawled by an information science Web crawler, designed for exhaustive elimination of duplicate pages (Thelwall, 2001a, 2001b). It recorded both the total number of HTML pages and the non-HTML pages on the site. The crawler would only visit

pages that could be found by following links in HTML pages, starting from the site home page. A commercial search engine crawler would potentially have two advantages in its ability to cover a site; it may know of other pages by (a) links to them from pages in other sites, and (b) having crawled the site before and remembered the locations of pages not currently linked. Consequently commercial search engines could theoretically find more pages than the research crawler. However, this would be true for all countries so the comparison among different countries in the study would not be negatively affected.

STEP 2: Manual filtering of sampled sites

Each Web site obtained in the above process was manually checked to filter out the following types of site that do not fit the definition of the study.
- Personal sites, e.g. sites containing baby or wedding photographs. These are not truly commercial in nature.
- Sites that explicitly indicate that they are under construction.
- Sites that are selling the domain name owned.
- Sites that are password protected and not accessible to search engines.
- Sites that are search engine interfaces themselves.

In addition, sites that do not fit the language requirements were filtered out. These included Chinese and Taiwanese sites that are not in the Chinese language and Singapore sites that are not in English. Very few sites in these countries do not fit this requirement. For example, among all the Singapore sites examined (142 in total), only one was in Chinese and one in another language.

Each of 330 .com sites selected by the computer program was manually checked to see if it represented a U.S. company. For the feasibility of data collection, a U.S. company is operationally defined as a company for which the main contact address on the Web site is in the U.S. This means that a U.S. subsidiary in a foreign country would not be considered a U.S. company, which is the way it should be. If a .com site was not from the U.S. but from one of the other countries in the study, it was put into the sample for that country. For example, among the 330 .com sites examined, 39 belonged to companies in China and they were moved into the China sample. This ensured that all Chinese companies, whether their URLs end with .com or .com.cn, had a chance to be included in the study.

Of the 330 .com sites examined, 143 (43%) were U.S. company sites. We kept examining sites from China and Taiwan sampled by the computer program until they reached a similar number, 143 sites from China and 141 sites from Taiwan. After examining the entire Singapore sample (142 sites) generated by the computer program, there were 94 that fitted the requirement of the study. In summary, the sample sizes for the U.S., China, Taiwan, and Singapore were 143, 143, 141, and 94 respectively, i.e. a total of 521 sites were included in the study.

Determining Site Coverage by Search Engines

Three major commercial search engines, Google, AltaVista, and AllTheWeb, were selected for the study. These three engines were among the top engines by size at the time of the study (Notess, 2002). Each of these engines can report the number of pages indexed for a particular site. The syntax of the search query to determine site coverage by the three engines will be explained using an imaginary site www.abc.com. For AltaVista, the search was done in the advanced search mode by the query "host:www.abc.com". The search was carried out in the advanced search mode of AllTheWeb by entering "www.abc.com" in the "domain filter"

window. Google has the "site" command to restrict a search to a particular Web site. However, this search command cannot be used alone and must be used in conjunction with another search command through a Boolean relationship. For example, search query "business site: www.abc.com" will retrieve all pages that contain the word "business" on the site www.abc.com. By reasoning, "-jfldsa site: www.abc.com" will retrieve all pages on the site www.abc.com that do not contain the word "jfldsa" (the minus sign in front of a word means not to contain the word"). Since "jfldsa" is not a valid word, this command will effectively retrieve all pages on the site. This is the strategy we used for the site coverage searches in Google.

It is known that the technical features of pages can have an impact on the ability of crawlers to index them (Thelwall, 2000). For example, pages with links in Flash, Java or JavaScript instead of directly in the HTML would not be indexed properly. As a result, sites using these features may be less well covered than others using standard HTML. In our study, this should not be a problem since in all cases crawlers were used to visit and index the sites (either our own crawler or that of a search engine) and the comparison of the two crawlers' data was the determining factor, so there should be no bias for sites with any unusual design feature.

Searching for Links to the Web Sites

All three search engines have the ability to search for links to a particular Web site (also referred to as "inlinks"). AltaVista and AllTheWeb can further restrict the search to external links: links coming from sites other than the one in question. However, Google does not have this capability. Since only external links will help the search engine to "see" the site in question (if the site is not covered by the search engine, the engine will not "see" the site no matter how many links there are among pages within the site), an external link search is need to investigate the question of whether site visibility could affect the coverage of the site by a search engine. Therefore, Google had to be excluded from this investigation. The imaginary site of www.abc.com will be used to explain the syntax of the query to search for external links in AltaVista and AllTheWeb. In AltaVista, the command line "link:www.abc.com AND NOT host:www.abc.com" was entered into the query window of the advanced search mode. In AllTheWeb's advanced search mode, "www.abc.com" was entered into the window in between the windows of "must include" and "in the link to URL" under "Word Filters". Further, "www.abc.com" was entered in the "exclude" window under "Domain Filters".

Search engine results have been very changeable over time in the past with the same query giving different results over short periods of time (Bar-Ilan, 1999; Rousseau, 1999; Mettrop & Nieuwenhuysen, 2001), although the instability seems to have reduced recently (Thelwall, 2001c; Vaughan & Thelwall, 2003). A study on search engine performance evaluation (Vaughan, 2003) conducted in the summer of 2002, around the time of data collection for this project, found that search engine output to be much more stable than what is reported in Selberg and Etzioni (2000). Two of the three engines in this study (Goole and AltaVista) were covered in Vaughan (2003) study. Nevertheless, search engine stability issue was taken into consideration in the current study. To improve the reliability of data collected and ease out the possible volatility of search engine performance, two rounds of data were collected for the same link search query with about three weeks in between the rounds. Results from both search engines were found to be very stable in that the two rounds of data were highly correlated (Spearman correlation coefficient is 0.999 and 0.992 for AltaVista and AllTheWeb respectively. In fact, the two rounds of AltaVista link counts were identical for 90% of sites). The average of the two rounds of data were taken and used in all the statistical analyses involving this variable.

Data Analysis and Results

Search engines' coverage of sites for different countries was examined in two ways: the percent of sites covered and the site coverage ratio.

Percentage of sites covered

The percentage of sites covered is summarized in Table 1. The average coverage with all countries and search engines combined is about 61%. It is clear that Google covers more, followed by AllTheWeb. AltaVista has the least coverage. The breakdown by country shows a very strong contrast between the U.S. and other countries. It is obvious that the discrepancy cannot be simply attributed to the language factor as the U.S. and Singapore both use English and yet U.S. coverage is much higher. China and Taiwan both use the Chinese language and their coverage is actually better than that of Singapore, other than the fact that AltaVista covers only 4% of Taiwanese sites. The extremely low coverage of Taiwanese sites by AltaVista was very surprising and we are not aware of any particular reason for this. The cause of this low coverage could not be technical because Google and AllTheWeb both have fairly good coverage of Taiwan, which in fact is better than their coverage of China.

Table 1 Percentage of Web Sites Covered

	U.S.	China	Singapore	Taiwan	Average
Google	87%	70%	56%	75%	72%
AllTheWeb	83%	61%	50%	75%	67%
AltaVista	80%	52%	41%	4%	44%
Average	83%	61%	49%	51%	61%

Site Coverage Ratio

For each site in the study, the site coverage ratio was calculated as follows:

$$\frac{\text{Number of pages covered by the search engine}}{\text{Number of pages found by the research crawler}}$$

For each site visited, the crawler recorded both the number of HTML pages and the total number of pages in the site. Because AltaVista and AllTheWeb index only HTML pages whilst Google indexed various types of pages at the time of data collection, (Notess, 2002), the site coverage ratio was calculated accordingly. The denominator of the "site coverage" formula is the total number of pages for Google and the number of HTML pages for AltaVista and AllTheWeb. Each site has three coverage ratio figures for the three engines respectively. When data for the three search engines were combined, the median coverage ratios for U.S. China, Taiwan, and Singapore are 0.89, 0.22, 0.03, and 0 respectively. This means that typically 89% of pages on a

U.S. site were covered. In contrast, only 22% of pages from China and 3% of pages from Taiwan were covered.

A two-way analysis of variance (ANOVA) test was carried out to further examine the coverage ratio by country and by search engine. The dependent variable of the test is the coverage ratio and the two independent variables are country and search engine. The frequency distribution of the dependent variable is very skewed, which is not surprising given the ubiquity of power-law type phenomena on the Web (Rousseau, 1997; Broder *et al*., 2000; Thelwall & Wilkinson, 2003). However, this violates the normality requirement of the ANOVA test. A logarithmic transformation is needed (Judd & McClelland, 1989, 493-528; Howell, 2002, 342-349). The frequency distribution is no longer skewed after the logarithmic transformation and an ANOVA test was carried out. The test result shows that there is a highly significant difference in coverage ratio (p<0.001) among different countries and different search engines. There is also a significant interaction between these two independent variables (p<0.001). Figure 1 summarizes the coverage ratio data.
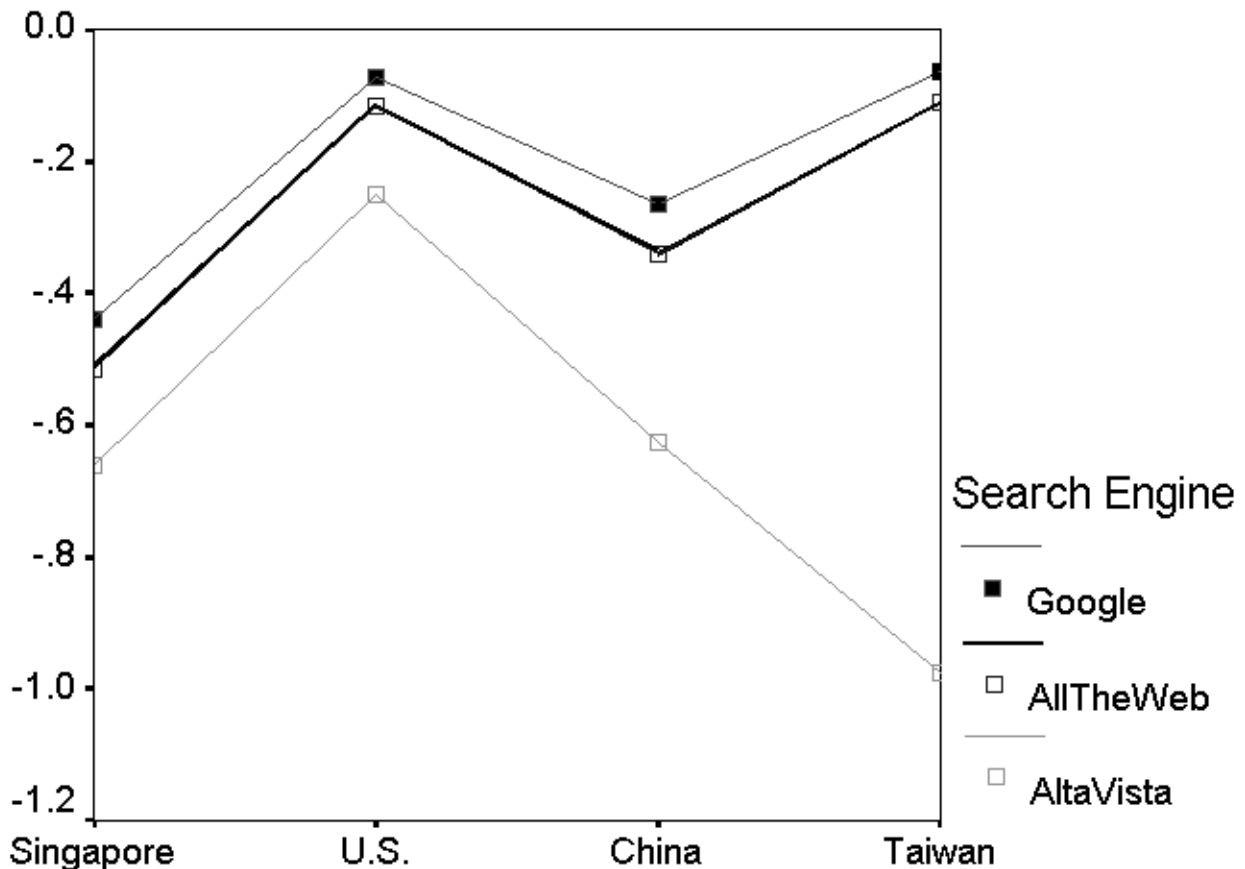

Figure 1 Coverage Ratio Comparison

The vertical axis represents the mean of the logarithm of the coverage ratios (coverage ratios are under 1 and become negative after the logarithm transformation). The horizontal axis

represents countries. The three lines on the graph represent the three search engines under comparison. It is clear that Google's coverage ratio is higher than that of AltaVista and AllTheWeb (higher on the graph means higher coverage ratio). U.S. sites are better covered on average and this cannot be explained by an advantage of the English language. Site language does not seem to be factor here because Singapore (English) fairs no better than China and Taiwan (Chinese). The three lines are not parallel, which means that there is an interaction between the two independent variables: country and search engine. The graph shows the pattern of the interaction: Google and AllTheWeb cover Taiwan more than China and Singapore whilest AltaVista does the opposite. In fact, Google and AllTheWeb have similar pattern of coverage for all four countries.

The Correlation between Coverage and Links to Web Sites

As discussed in the methodology section of the paper, links to a Web site were investigated as a possible cause of the biased coverage of different countries. The hypothesis is that the coverage of a site is correlated with the number of links to the site. The Spearman correlation coefficient was used to test this hypothesis (the frequency distributions for both the link counts and the coverage ratio were very skewed so the Pearson correlation test is not appropriate). The correlation test was performed separately for the different search engines. In other words, the number of links found by a particular search engine was correlated with the coverage ratio of that engine. The reason for doing this is: if sites that link to a particular site are not indexed by the search engine (i.e. the link search using this search engine results in zero hits), the engine will not know the existence of that site even if these links are found by another search engine. The correlation was highly significant ($p<0.001$) for AllTheWeb and AltaVista; 0.6 and 0.47 respectively (the test was not performed for Google because no link data were collected from this search engine as explained in "Searching for Links to the Web Sites" section of the paper). We consider both correlations to be fairly strong given the sample size (521 data points).

Comparing Coverage with Link Counts Taken into Consideration

Since the coverage ratio was found to correlate with the links to a site, it is desirable to repeat the coverage analysis with link counts taken into consideration. No previous study has documented how this should be achieved, however. In particular, how should the number of links to a site affect its chance of being indexed by a search engine? In theory, once a search engine has found *one* link to the site, it should be able to index it. AltaVista insiders (Broder *et al.*, 2000) have hinted at five links as an alternative cut-off point for a page to be indexed so other minimum values are also possible. Alternatively, larger inlink counts may push a new site higher up the list of new sites to crawl, or the priority to crawl could also correlate with site age with older sites presumably more likely to have already climbed near the top of the waiting list, assuming that there is such a list for a search engine. In the light of the lack of sufficient knowledge and evidence, we will fit a simple model to the data.

A new variable, *site coverage per link*, was created and calculated for each site. This is defined to be the coverage ratio divided by the number of links to the site. However, this new variable applies only to sites that have links to them and Google is excluded from this analysis since no link data was collected for Google. The frequency distributions for the site coverage per link variable were very skewed so the logarithmic transformation was again applied. A two-way analysis of variance test was then carried out with country and search engine as the independent

variables; a similar analysis to that carried out for the coverage ratio data discussed above. The test shows that there is a significant difference (p<0.05) among different countries and different search engines. There is also a significant interaction between these two variables (p<0.001). Figure 2 summarizes the site coverage per inlink comparison in the same way as that for Figure 1.
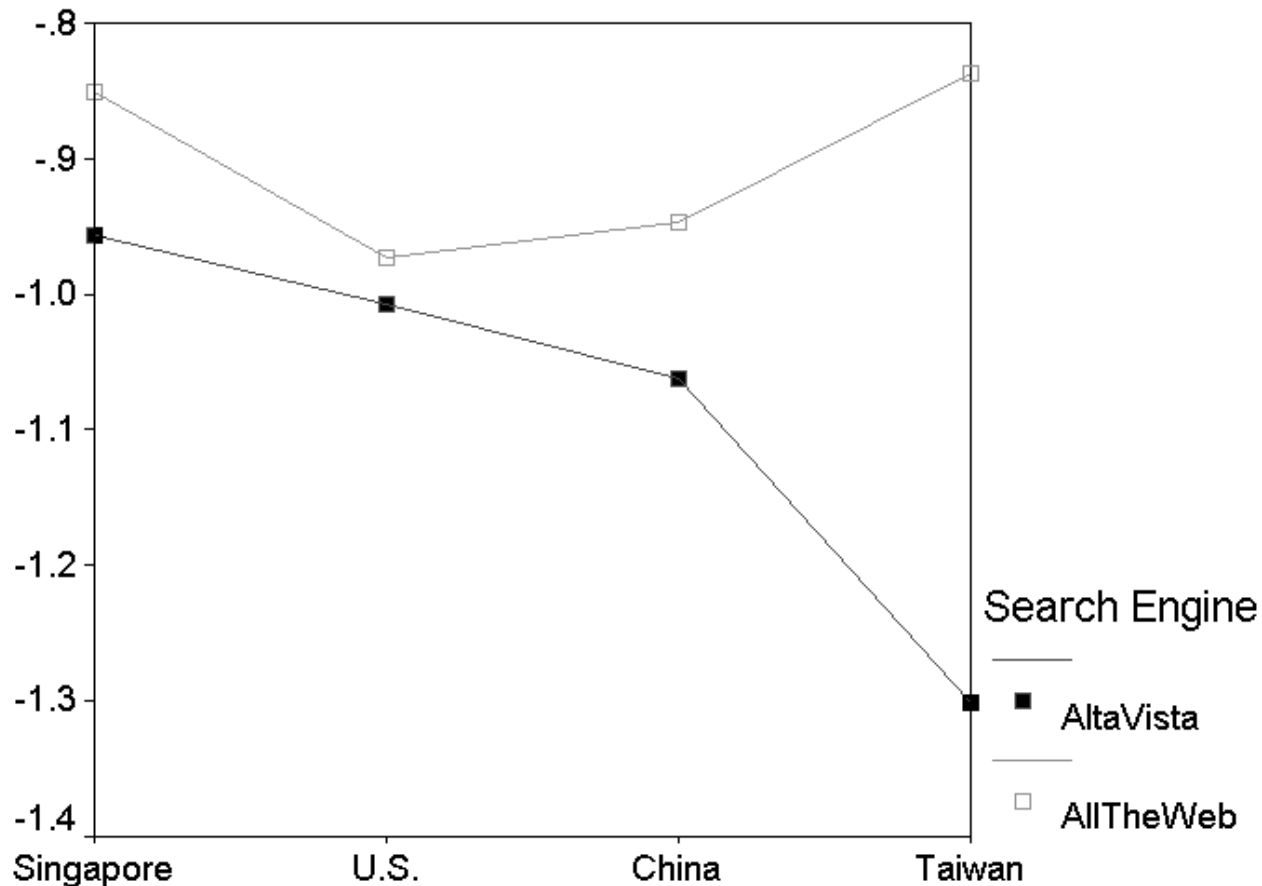

Figure 2 Site Coverage per Link Comparison

It is clear that AllTheWeb records higher site coverage per inlink than AltaVista. This echoes the conclusion reached in the coverage ratio analysis reported above. However, U.S. coverage is no longer above other countries on average as was the case in Figure 1. In fact, the U.S. coverage is lower than that of Singapore and similar to that of China. Again, language does not seem to be a factor here. The two lines on Figure 2 are not parallel which shows different national trends for the two search engines. AllTheWeb covers Taiwan better (on a per link basis) than China, which in turn is more than the U.S., while AltaVista works in the opposite direction for these countries.

Discussion

Search engines do not cover all of the Web sites available, or even all of the Web sites that they know about from the links in their own databases. On average, about 61% of sites in the study were indexed by the three search engines. Overall, Google covers most sites followed by AllTheWeb, and AltaVista has the lowest coverage. Even when a site was indexed, not all of its

pages on the site were indexed. The coverage ratio, calculated as the percentage of pages on a site that were indexed by a search engine, is less than one on average. This means that the research crawler found more pages than were indexed by the search engines. When coverage of different countries was compared, a very strong pattern of uneven coverage emerges: U.S. sites get much higher coverage than those of China, Taiwan and Singapore. This is true whether the coverage is measured by the percentage of sites covered or the percentage of pages on a site that are indexed. Typically 89% of pages on a U.S. site were covered. In contrast, only 22% of pages from China and 3% of pages from Taiwan were covered. Over 50% of the *sites* from Singapore were not covered at all. The national variations found may not be taken literally as shown by these numbers because our sampling method was relatively random rather than absolutely random, although we see no significant bias in the sample. Nevertheless, we claim that the differences found are compelling enough to make the conclusion justifiable.

Two possible technical causes for the biased coverage were explored in the study: the language and the visibility of the site. The language factor was examined by comparing the English sites (U.S. and Singapore) against the Chinese sites (China and Taiwan). In all the comparisons made, language does not seem to be a factor that affects the chance of a site being indexed. In other words, the possible technical difficulty of processing the Chinese language does not seem to deter search engines from covering sites with the Chinese language.

The visibility of a site is measured by the number of links from other sites to the site in question. A statistically significant correlation was found between link counts to a site and the site's coverage by search engines. In general, the more links a site receives from other sites indexed by the search engine, the higher the proportion of the site that was covered by the search engine. This does not necessarily mean that the more links a site attracts from *all* other sites, the more coverage it will get from the search engine, because of the large number of sites that are not indexed by search engines. This is actually a critical point that we will analyse further. The Web is a dynamic rather than static environment, so time is a factor in site indexing. Since the U.S. had an early start, its Web sites have had longer to be found by search engines. Previous studies have shown that sites tend to link more within their own country than outside (Bharat et al, 2001; Thelwall, 2002b). Since at any given point in time a higher proportion of links to U.S. sites is likely to be indexed by search engines, even the new sites in the U.S. are more likely to be indexed than the new ones in other countries. This is a result of any sites that link to them being more likely to be indexed, through being more likely to be in the U.S. Essentially, then, the larger the percentage of sites in a country that a search engine has indexed, the more likely it is that new sites from the country are found and indexed, a success-breeds-success effect. This is broadly consistent with our findings. There is a possible counteracting tendency for low Web using countries to attract a disproportionate share of international links (Thelwall & Smith, 2002), which may serve to give them an initial boost. Logically, however, in the long term if the Web continues to increase exponentially in all countries then national differences in coverage will continue. Conversely, if web page creation starts to reach saturation level, which does not seem to be the case currently, then it could be expected that the current national differences should naturally wither as search engines gain the ability to index all pages that they can find links to.

Since the links to a site were found to affect site coverage, the coverage comparison among different countries was extended by taking link counts into consideration through a simple coverage per link model, the coverage ratio divided by the inlink count. The analysis of this model showed a picture very different from the one based on coverage alone. The U.S. was no longer the best-covered country when inlinks to Web sites were taken into consideration. This

confirms that the higher coverage of U.S. sites could be simply the result of more links to those sites being indexed.

Conclusions

First, the lack of linguistic problems for site indexing is highly reassuring, although this is only the indexing side of Web information retrieval. The other side is search queries. Current evidence suggests that there are still outstanding issues for multilingual text querying (Moukdad, 2002). Differentiated site coverage was found for the four countries in the study. The evidence is consistent with site link counts being a major contributory factor, and the unequal coverage being a consequence of the combination of this factor with the exponential growth of the Web, the early start of the US in Web creation and the possible tendency for sites to link to others in the same country.

The biased coverage in favour of U.S. sites appears to be caused by technical reasons. It should be acknowledged, however, that the study made no attempt to investigate political or social causes, which are outside the scope of an empirical study like this, so no firm conclusions can be drawn in that regard. The likely continued imbalance is still a political issue, however, and there are grounds for countries outside the U.S. to make an argument for the economic necessity of affirmative action by major search engines to even out their coverage. Our findings indicate that any such discussion should not start with the premise of the existence of a deliberate bias, only an unintentional bias resulting from the success-breed-success or cumulative advantage effect. Of course, the market force solution to this problem is simultaneously operating: national and regional search engines that can presumably gain much higher coverage of their chosen area are being developed and used.

The results of the study re-emphasise the importance of Web visibility for businesses. A business Web site should try to attract as many links from other sites as possible, particularly other sites that are *already indexed by search engines*, irrespective of national origin. Our study therefore confirms Walker's (2002) assertion that links have economic power. Extra links will not only directly increase the traffic to the site but also the likelihood for the site being indexed by search engines, generating additional potential visitors.

As an initial quantitative analysis of search engine coverage bias, the study only investigated three search engines (major search engines though), four countries, and two languages. The conclusions reached are all limited to this context. Further research involving more search engines and countries are warranted and may lead to different conclusions. This is definitely an area worth pursing due to the increasing importance of search engines and the Web in general.

References

Albitz, P. & Liu, C. (2001). *DNS and BIND (4th ed.)*. Sebastopol, California: O'Reilly.

Bar-Ilan, J. (1999). Search engine results over time - A case study on search engine stability. *Cybermetrics*, 2/3. Available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

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

BBC (2002). China criticised for ban on Google. Retrieved Nov. 27, 2002 from http://news.bbc.co.uk/1/hi/technology/2238236.stm.

Bharat, K., Chang, B., Henzinger, M., & Ruhl, M. (2001). Who links to whom: mining linkage between Web sites, *Proceedings IEEE International Conference on Data Mining (ICDM)*,

San Jose, Nov. 2001. Retrieved June 9, 2003 from
http://theory.lcs.mit.edu/~ruhl/papers/2001-icdm.pdf.

Broder, A. Kumar, R, Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. &
Wiener, J. (2000). Graph structure in the Web, *Journal of Computer Networks*, 33(1-6),
309-320.

Gordon, M. & Pathak, P. (1999). Finding information on the world wide web: the retrieval
effectiveness of search engines. *Information Processing & Management*, 35(2), 141-180.

Grefenstette, G., & Nioche, J. (2000). Estimation of English and non-English language use on the
WWW. *Proceedings of the RIAO'2000 Conference*. Paris: C.I.D. Retrieved Dec. 20, 2001,
from http://133.23.229.11/~ysuzuki/Proceedingsall/RIAO2000/Wednesday/20plenary2.pdf

Hawking, D., Craswell, N., Bailey, P., & Griffiths, K. (2001). Measuring search engine quality.
Information Retrieval, 4(1), 33-59.

Henzinger, M. R., Heydon, A., Mitzenmacher M. & Najork, M. (1999). Measuring index quality
using random walks on the Web, *Computer Networks and ISDN Systems*, 31(11-16), 1291-
1303.

Howell D. (2002). *Statistical Methods for Psychology*, 5th ed., Pacific Grove, CA, U.S.A.:
Duxbury.

Huberman, B. A. & Adamic, L, A. (1999). Growth dynamics of the world wide web. *Nature*,
401, 131.

Introna, L. D. & Nissenbaum, H. (2000). Shaping the Web: why the politics of search engines
matters. *The Information Society*, 16, 169-185.

Ju-Pak, K. H. (1999). Content dimensions of web advertising: A cross-national comparison.
International Journal of Advertising, 18(2), 207-231.

Judd, C. M. & McClelland, G. H. (1989). *Data Analysis: A Model-Comparison Approach*, San
Diego, U.S.A.: Harcourt Brace Jovanovich.

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

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

Moukdad, H. (2002). Language-based retrieval of web documents: an analysis of Arabic-
recognition capabilities of two major search engines, *Proceedings of the 65th ASIST Annual
Meeting Volume 39 (ASIST 2002)*, pp. 551.

Mowshowitz, A. & Kawaguchi, A. (2002). Assessing bias in search engines, *Information
Processing & Management*, 38(1), 141-156.

Nie, J. Y. & Ren, F. (1999). Chinese information retrieval: Using characters or words?
Information Processing & Management, 35(4), 443-462.

Notess, G. (2002). Search Engine Statistics: Relative Size Showdown. Retrieved Aug. 1, 2002
from http://www.searchengineshowdown.com/stats/size.shtml.

OCLC Web Characterization Project (2002). Country and Language. Retrieved Sep. 4, 2002 from
http://wcp.oclc.org/

O'Neill, E. T., McClain, P. D. & Lavoie, B. F. (1997). A Methodology for Sampling the World
Wide Web, Retrieved Dec. 2, 2002 from
http://www.oclc.org/research/publications/arr/1997/oneill/o%27neillar980213.htm.

Oyama, K., Ishida, E. & Kando, N. (2003): *NTCIR Workshop3: Proceedings of the Third NTCIR
Workshop on research in information Retrieval, Automatic Text Summarization and*

Question Answering. Retrieved June 9, 2003 from
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html.

Peters, C. (Ed.) (2001). *Cross-language information retrieval and evaluation: workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000: revised papers*. Lecture Notes in Computer Science, 2069. Berlin: Springer.

Rousseau, R. (1997). Sitations: an exploratory study, *Cybermetrics*, 1. Retrieved Dec. 2, 2002 from http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight, *Cybermetrics*, 2/3. Retrieved Dec. 2, 2002 from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Selberg, E. & Etzioni, O. (2000). On the instability of Web search engines. *Proceedings RIAO*, Paris. April 2000. PDF file retrieved June 9, 2003 from http://citeseer.nj.nec.com/selberg00instability.html.

Sullivan, D. (2001). AltaVista regional listings left to rot. Retrieved Nov. 27, 2002 from http://searchenginewatch.com/sereport/01/09-altavista.html.

Thelwall, M. & Smith, A. (2002). A study of the interlinking between Asia-Pacific university Web sites, *Scientometrics* 55(3), 363-376.

Thelwall, M. & Wilkinson, D. (2003). Graph structure in three national academic Webs: Power laws with anomalies, *Journal of the American Society for Information Science and Technology*, 54(8), 706-712.

Thelwall, M. (2000). Commercial Web sites: lost in cyberspace?, *Internet Research: Electronic Networking and Applications*, 10(2), 150-159.

Thelwall, M. (2001a). A web crawler design for data mining, *Journal of Information Science* 27(5) 319-325.

Thelwall, M. (2001b). A publicly accessible database of UK university website links and a discussion of the need for human intervention in web crawling, Retrieved Jan. 14, 2003 from http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf

Thelwall, M. (2001c), The responsiveness of search engine indexes, *Cybermetrics*, 5(1), Retrieved Jan. 14, 2003 from http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html.

Thelwall, M. (2002a). Methodologies for crawler based Web surveys, *Internet Research: Electronic Networking and Applications*, 12(2), 124-138.

Thelwall, M. (2002b). Evidence for the existence of geographic trends in university web site interlinking, *Journal of Documentation*, 58(5), 563-574.

Vaughan, L. (2003). New measurements for search engine evaluation proposed and tested. To appear in *Information Processing & Management*.

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

Walker, J. (2002). Links and power: the political economy of linking on the Web. In: *Proceedings of ACM Hypertext 2002*, 72-73.