informationresearch

VOL. 14 NO. 2, JUNE, 2009

The effectiveness of Web search engines to index new sites from different countries

Ari Pirkola
Department of Information Studies, University of Tampere, 33014 Tampere, Finland

Abstract

Introduction. *Investigates how effectively Web search engines index new sites from different countries. The primary interest is whether new sites are indexed equally or whether search engines are biased towards certain countries. If major search engines show biased coverage it can be considered a significant economic and political problem because of the international nature of the major search engines.*
Method. *We examine what share of the sites of recently registered domain names from a certain country appears in a search engine index after a given period of time following registration of the domain name. We consider how effectively the Websites of new Finnish, French U.S. domain names are indexed by two US-based major search engines (Google and Microsoft's Live Search) and three European search engines (Virgilio, www.fi Voila).*
Results. *The results showed that Google provided the highest coverage of the five search engines that US-based search engines Google and Live Search indexed US sites more effectively than Finnish and French sites. These findings are in line with earlier research findings based on a different method and different countries. The Finnish www.fi indexed only Finnish sites and the French Voila only French sites. Virgilio indexed European sites more effectively than US sites.*
Conclusions. *The biased coverage of Google and Live Search raises concern because of their international nature. The coverage bias by the European search engines only seems to have local or regional significance.*

CHANGE FONT

Introduction

Currently the World Wide Web contains billions of publicly available pages. Besides its huge size, the Web is characterized by its rapid growth and rate of change. A vast number of new sites and pages are created every day. As more information becomes available on the Web it is more difficult to provide

effective search services for Internet users. Web search engines, such as Google and Microsoft's Live Search, provide access to indexable Web documents (pages), but because of the Web's immense size each search engine is able to index only a portion of the entire indexable Web (Barfourosh *et al.* 2002, Castillo 2004). Therefore. a vast amount of information, maybe billions of Web documents, is hidden from Internet users. Because of the limited site and page coverage a search engine may be biased to certain countries. The global search engine market and access to the Internet content is dominated by US-based commercial search engine giants there is empirical evidence that US-based search engines favour U.S. Websites (Vaughan and Thelwall 2004, Vaughan and Zhang 2007).

Proportionally smaller coverage of certain types of Websites in search engines, for example, sites of certain countries, results in the decreased visibility of those sites on the Web. Because of the significance of the Web as a source of information in today's world and the international nature of the major search engines the decreased visibility can be considered a significant economic and political problem (Van Couvering 2004, Vaughan and Zhang 2007). A company whose site is not included in the database of a search engine may experience a decline in revenue. If sites are not indexed by search engines Internet users may lose important health related information, product information, education material other useful information sources.

An individual or organization publishing a Website has to acquire a domain name for the site, a unique alphabetical address, e.g. *www.microsoft.com* has to register it. The registration is provided by Web hosts, which hat also provide server disk space for their clients for storing and maintaining the sites. In this study, we investigate how effectively Web search engines index new sites from different countries. We examine what share of the sites of recently registered domain names from a certain country appears in a search engine index after a given period of time following the domain name registration. Site coverage is considered from the European viewpoint and we are interested in how effectively the Websites of new Finnish, French U.S. domain names are covered (indexed) by US-based and European search engines. Being the home country of major search engines, the U.S. serves as a reference country: search engine coverage of new Finnish and French sites is compared to search engine coverage of new U.S sites.

Information contained in new Websites can be considered to be particularly valuable for many Internet users the question of the new site coverage of search engines as such is an interesting and important research problem. However, the present study considers search engine coverage also from a more general perspective, since we follow the increase of the coverage up to half a year after the registration of the domain names of the sites.

For each of the three countries, recently registered domain names were taken from domain name sources (e.g. the Ficora domain name registry). After eleven and twenty-five weeks of the registration, the active sites of the domain names were searched for using two major US-based search engines (*Google* and *Live Search*), a large European search engine (*Virgilio*) two country-specific search engines (Finnish *www.fi* (after this study was completed the Finnish search service www.fi was reorganised and renamed 02.fi Fonecta) and French *Voila*). The analysis of the achieved data allows us to answer the following research questions: (1) Which of the examined search engines achieves the best coverage rate? (2) Are new sites from different countries indexed equally? If not, towards which countries are different search engines biased? (3) How quickly are the sites of new domain names indexed by the search engines?

In this study, we take the same approach to search engine coverage as Vaughan and Zhang (2007). Regarding global search engines (Google and Live Search), an ideal situation would be that a search engine would cover the same proportion of Websites from different countries, i.e., sites from different countries would have an equal chance of being indexed. In contrast to this, a country-specific search engine is expected to mainly index the sites of that country.

Related Work

It seems impossible to determine the exact size of the Web and the coverage of different search engines. There are, however, estimates of them. In 1999 it was estimated that no search engine indexed more than 16% of the indexable Web (Lawrence and Giles 1999). The size of the indexable Web was reported

to be 800 million pages. The study by Gulli and Signorini (2005) estimated that, as of January 2005, the indexable Web covered approximately 11.5 billion pages and that Google's coverage rate was 76.2%. For MSN and Yahoo! the estimated coverage rates were 61.9% and 69.3%. (It should be noted that the above figures only refer to Web page coverage, not to site coverage.) Unfortunately, the study by by Gulli and Signorini (2005) says nothing about the reliability of the reported figures. Nevertheless, it seems clear that search engine coverage has increased considerably from what it was in the late 1990s.

As shown above, the Web consists of a vast number of documents. It is also characterized by its rapid change rate. The study by Ntoulas *et al.* (1999) illustrates this point. The researchers measured the change in the Web's content and link structure from the viewpoint of designing effective search engines. Representative snapshots of Websites were collected during a one year period. Based on their experimental results, the researchers estimated that only 40% of Web pages of today will still be accessible after one year and that 640 million new pages are created every week. The most dramatic changes appear in the link structure of the Web: around 80% of all links are replaced within a year.

Such a rapid change implies that search engines often provide users with outdated information. Lewandowski (2004) and Lewandowski *et al.* (2006) investigated the ability of three major search engines (Google, Teoma and Yahoo!) to retrieve recent versions of documents. Both studies showed that the tested search engines did not perform satisfactorily in this regard. In Lewandowski (2004) even the best search engine, Google, did not return more than 60% of the documents that were updated within a period of six months before the retrieval experiments.

Limited site and page coverage of search engines is related to a coverage bias. Empirical research has shown that major US-based search engines favour U.S. Websites (Thelwall 2000, Vaughan and Thelwall 2004, Vaughan and Zhang 2007) Thelwall (2000) compared search engine coverage of some 60,000 sites from forty-two different countries (domains). The tested search engines were AltaVista, Hotbot, InfoSeek and MSN Yahoo! The study showed that some countries received consistently higher coverage rates than some other countries across the five search engines. For example, Altavista and MSN covered 82.0% and 71.0% of the .com sites (presumably most of them were U.S. sites), but only 37.0% and 25.0% of the Egyptian sites.

Vaughan and Thelwall (2004) studied country biases in the coverage of three main search engines (AllTheWeb, Altavista and Google) using randomly generated domain names as the test data. The percentage of commercial sites found by a research crawler not dependent on the search engines was first determined. Then we examined what share of these sites the search engines returned. The study found significant differences in the coverage: the search engines indexed a considerably larger proportion of U.S. sites than sites from China, Taiwan and Singapore. These results were confirmed in Vaughan and Zhang (2007) who found that major search engines (e.g., Google) indexed U.S. commercial sites more effectively than commercial sites from China, Taiwan and Singapore. Also, the average coverage of governmental, educational, organizational and commercial sites was better for the U.S sites than for the sites of the three other countries.

There is a concern among researchers about the hegemony of US-based search engines because of the economic and political aspects and the worldwide significance of the search engines (Introna and Nissenbaum 2000, Van Couvering 2004, Vaughan and Thelwall 2004, Vaughan and Zhang 2007). The results reported in this study support the issues raised in the literature.

Mowshowitz and Kawaguchi (2005) proposed a measure of bias for evaluating performance differences between search engines and they showed that the performance of search engines can be distinguished by means of the proposed measure. The measure compares the results of one search engine against those of a control group. In the present study, bias refers to a situation where Websites from different countries are not indexed equally by (major) search engines, rather than to an average based on a set of search engines. It is important to keep the two concepts distinct from each other.

The contribution of the present paper focuses on three issues. First, this is the first study to investigate how effectively and quickly Web search engines index new Websites. Like Thelwall (2000), Vaughan and Thelwall (2004) and Vaughan and Zhang (2007), we address the issue of search engine coverage and examine whether search engines favour or disfavour certain countries. The main difference is that

the present study considers the sites of recently registered domain names, whereas the above studies tested established Websites. The test data in these studies consisted of randomly generated domain names, whereas we systematically selected the new domain names from domain name sources. The tested countries were also different. Secondly, the above studies demonstrated that in the case of established Websites major US-based search engines are biased towards U.S. sites. In this study, we demonstrate that this holds also for the Websites of recently registered domain names. Third, we demonstrate that different types of search engines show great variations across different countries in the coverage of new Websites.

Methods and data

In this section, we first describe the selection of the Finnish, French U.S. domain name samples. Unlike the Finnish and French samples, U.S. samples were taken from a secondary domain name source to ensure that they represent all new U.S. domain names and the U.S. domain source was analysed extensively. The analysis is described in a separate subsection. In the last subsection, we consider the search engines and queries used in the experiments the evaluation of results.

Selection of the Finnish, French U.S. samples

First we describe the general approach applied in the selection of the test domain names then we describe the selection process in more detail. As test data we used new Finnish, French and U.S. Websites, i.e., the sites of new domain names with the extensions *.fi* (for Finnish), *.fr* (for French) *.com*, *.org .net* (for the U.S.). The sites of the domain names were searched for using the five search engines eleven and twenty-five weeks after the domain names were registered. The Finnish domain names were registered by commercial Finnish hosts the sites were located on servers that were located in Finland. Correspondingly, the French and U.S. domain names were registered by commercial French and North-American hosts the sites were located on French and U.S. servers. Two domain name samples were taken for each of the three countries. The domain names of the first three sets were taken in the spring 2007 and those of the second in the summer 2007. In this manner, variation over time was generated. In each six case, sampling consisted of three stages. In the first stage, a large set of recently registered domain names was taken from a domain name source. In the second stage, inactive sites were removed and only active sites were kept in the test data set. In the third stage, the locations of servers where the sites of the domain names were located were identified. The sites located on servers located in countries other than the country in question were removed from the test data. The location of every Nth site in the list produced in the second stage was checked iteratively, so that each final sample contained 200 systematically selected domain names.

The selection of the first Finnish sample is described next. The second Finnish sample, as well as the French and U.S. samples were chosen in a similar manner. There was, however, slight variation in the selection process which is described below.

The study started in May 2007. In the first stage, all domain names registered in May 10-24, 2007 were taken from the Ficora domain name registry. The registry provides lists of all Finnish (.fi) domain names, including all recently registered names. In the second stage, three weeks after registration of the domain names, we downloaded by a Web browser the pages pointed to by the domain names and reviewed which of the downloaded sites were active. A time period of three weeks was applied based on the observation that not many Website publishers construct their sites immediately after the domain name registration. Inactive sites were removed and only active sites were kept in the data set. A new site was considered active if it included one or more pages that contained information created by the publisher of the site.

In the third stage, the active sites that were located on servers not located in Finland were removed from the test data. For example, the sites with .fi country extension that were located on a Swedish server were removed. In this study we consider new sites hosted by commercial hosts in this stage also the sites hosted by other types of hosts (e.g. universities and state agencies) were removed. Most of the hosts were commercial hosts this step removed only a few sites. Also for French, the sites hosted by non-commercial hosts were removed, whereas the U.S. data only contained sites hosted by commercial hosts. The server information was obtained by means of the *Network.tools* service, which provides a numerical

IP address for an entered domain name, the name of the server on which the site is located, the home country of the server, as well as other information related to the entered domain name. There are several similar services to Network.tools on the Web. We tested a few of them and found only minor discrepancies regarding information on the home countries of servers. The server home country information provided by Network.tools was also consistent with the domain name extensions. For example, the sites with .fi extension typically were located on Finnish servers.

The second Finnish sample, as well as the two French and the two U.S. samples, was selected in a similar manner to the first Finnish sample. However, because France and the U.S are much bigger countries than Finland and have more registrations daily, for the French and U.S. samples we could select the new domain names from a shorter registration time period (see below the actual registration days). For French, the domain names were selected from the Afnic domain name registry. It reports all recently registered French (.fr) domain names. The U.S. domain names were selected from the *Daily Changes* list by Name Intelligence. Each day a list of the most active Web hosts is published. The Web hosts under the title *Top [n] most active name servers on [date] with new domains* were considered in this study. Each entry in the list represents one host and includes a link to another list containing the new (newly created, see next subsection) domain names that host has registered (deleted and transferred names are also reported). The *Daily Changes* list also contains hosts whose domain names are registered for other purposes than for constructing Websites. They were first removed by reviewing which hosts had active sites and only the remaining hosts with active sites were used to select the final hosts. To get variation over hosts, for each U.S. sample, the 200 domain names were selected from four hosts, 50 names from each host. Thus, there were 400 U.S. domain names registered by eight different hosts. Host selection was similar to site selection explained above, in that every Nth host was selected with N being an arbitrary number.

The registration periods for the second Finnish sample and for the French and U.S. samples are presented below. The names of the eight U.S. hosts are also shown.

- Finnish, the second sample: July 23-31, 2007
- French, the first sample: May 21 and 28, 2007
- French, the second sample: July 19-20, 2007
- U.S., the first sample: May 25, 2007 (DreamHost , iPowerWeb, Microsoft Office Live, Mdnsservice (Tucows))
- U.S., the second sample: July 20, 2007 (Netfirms, HostGator, BlueHost , 1and1)

The following hosts had the largest number of domain names in the Finnish and French data: Finnish: Nebula.fi, Neobitti.fi, Netsor.fi and Planeetta.net (Planeetta Internet); French: Gandi.net, NFrance.com, Ovh.net and Sites-acs.net (which appears to have been taken over by Nordnet.net).

After the selection of the final samples, we reviewed all six samples to ensure that the registration day information provided by Ficora and Afnic *Daily Changes* was correct. A set of domain names was entered in Internet's Whois domain name registries. No major discrepancies were noted in this test.

Analysis of the U.S. domain name source

The Ficora and Afnic domain name registries report all recently registered Finnish and French domain names samples from the databases of these two registries represent all new domain names with the extensions .fi and .fr. The new U.S. domain names contained in Name Intelligence's *Daily Changes* lists are instead extracted from domain name servers that the company regularly monitors (the role of Internet's name servers is to translate human-readable domain names into IP addresses). The extracted new domain names and statistics associated with them are published in the *Daily Changes* list. The list reports new domain names with the extensions .com, .net, .org, .info, .biz .us. The new domain names were *created* on the publication day of *Daily Changes* or 1-2 days earlier, so they are genuinely new domain names (our Whois searches described above also confirmed this). The information on Name Intelligence's Web page suggests that its database covers all or most of the domain names with these extensions. We reviewed several *Daily Changes* lists in spring 2007. Each list contained a remarkable number of new domain names. To get a more accurate picture, we performed a detailed analysis of the

information and statistics contained in the list, the purpose of which was to ensure that our U.S. samples were representative of all new U.S. domain names that the comparison of search engine coverage of new Finnish and French versus new U.S. sites was fair.

There are two important questions related to the representativeness of the samples: (1) Are all important U.S. Web hosts included in *Daily Changes*? (2) Are all or most of the new domain names of the host listed?. To answer these questions, we compared host information and domain name statistics reported in the list to the information and statistics provided by two other name server monitoring companies: *WebHosting.Info* and *Ipwalk.com*. *WebHosting.Info* has an extensive name server monitoring system. The profile of the company is available, as is Ipwalk's methodology. Ipwalk's page says, for example, that it produces the statistics of domain name changes and growth based on advanced name server monitoring technology.

We acquired *WebHosting.Info*'s report on the U.S. hosts (n=20), which holds the largest number of domain names. *Ipwalk.com* has produced similar statistics that report the top ten U.S. hosts during the period October 2005 to September 2006. We also had available thirty-eight different *Daily Changes* lists from the years 2006 and 2007. All of them were used in this analysis, because all hosts monitored by Name Intelligence do not always appear in *Daily Changes* - this depends on the daily activity of the host. We reviewed the thirty-eight *Daily Changes* lists to find out whether the top U.S. hosts are included in *Daily Changes*. The results showed that out of the twenty hosts in *WebHosting.Info* 's report only one (Networksolutions.com) was not included in any of the *Daily Changes* lists (possibly, it may have had a different name in *Daily Changes*). All ten hosts listed in Ipwalk.com's report were found in *Daily Changes* (as well as in *WebHosting.Info* 's report). *Daily Changes*'s statistics were in line to the host rankings in the two reports. For example, *Wildwestdomains.com* held the largest number of domain names in all thirty-eight *Daily Changes* lists (the company is called *Secureserver.net* in *Daily Changes* according to its name server): it was ranked first in both reports. This first analysis showed that all (or nearly all) important U.S. hosts are included in *Daily Changes* that at least in this respect our U.S. samples are representative.

Daily Changes reports for all hosts contained in it the total number of domain names the host holds on the publication day of *Daily Changes*, i.e., the total number of domain names Name Intelligence has tracked. *Ipwalk.com* and *WebHosting.Info* have produced similar statistics. In the second analysis, *Daily Changes*'s total of domain names was compared to the statistics produced by *Ipwalk.com* and *WebHosting.Info*. All eight U.S. hosts presented previously were considered in this analysis. The rationale for this comparison was that if different name server monitoring companies report comparable figures and similar trends within the same period of time, this is a compelling evidence that *Daily Changes* contains all or most of the new U.S. domain names. We were in particular interested in whether the figures reported by *Daily Changes*, *Ipwalk.com* and *WebHosting.Info* match in magnitude. We also expected that each host holds more than 100 000 domain names. It is common knowledge to those working in the field that an average U.S. host holds several hundreds of thousands of domain names.

For *Daily Changes* and *Ipwalk.com*, the totals of domain names were recorded on June 7, 2007 and December 29, 2007. We did not have access to *WebHosting.Info*'s domain name statistics in June (if such exists), only the December statistics (December 31, 2007).

The results of the second analysis are presented in Tables 1 and 2. Table 2 shows the differences between December and June totals for *Daily Changes* and Ipwalk.com. The difference does not directly show the number of new domain names, but it reflects the number of new, deleted transferred domain names.

From Table 1 it can be seen that in all cases the totals reported by *Daily Changes*, *Ipwalk.com* *WebHosting.Info* are of the same magnitude. Table 2 in turn shows that the differences between December and June totals show the same trends for *Daily Changes* and *Ipwalk.com*. However, *Daily Changes*'s figures are systematically higher (except in one case where the domain name gain is negative); the figures show that *Daily Changes*'s coverage increase was higher in the latter half of 2007.

Based on the results of the two analyses we are confident that our U.S. domain names samples are representative of all new U.S. domain names.

Web host	Daily Changes June 7, 2007	Ipwalk June 7, 2007	Daily Changes December 29, 2007	Ipwalk December 29, 2007	Webhosting December 31, 2007
Dreamhost	509,329	521,788	629,250	624,376	634,923
Ipowerweb	422,033	431,502	374,566	397,195	406,953
Officelive	362,438	380,156	507,426	508,735	*
Netfirms	181,779	185,524	214,708	214,135	214,336
Hostgator	206,241	217,138	299,861	298,593	306,138
Bluehost	264,827	271,573	378,699	373,167	374,328
1and1	1,786,232	1,785,968	2,155,574	2,102,241	*
Mdnsservice	260,512	271,409	326,740	326,864	327,244

Table 1: Total number of domain names. (*Statistics not available.)

Web host	*Daily Changes* December-June	Ipwalk December-June
Dreamhost	119,921	102,588
Ipowerweb	-47,467	-34,307
Officelive	144,988	128,579
Netfirms	32,929	28,611
Hostgator	93,620	81,455
Bluehost	113,872	101,594
1and1	369,342	316,273
Mdnsservice	66,228	55,455

Table 2: Difference in total number of domain names.

Searching: search engine evaluation

After eleven and twenty-five weeks of the registration days, all the examined search engines, i.e., Google, Live Search, *Virgilio*, www.fi Voila were queried to see which sites of the test domain names each search engine had indexed. Furthermore, Google was considered in more detail in the beginning of the experimental part of the study using the May data: it was queried at weeks three, seven, eleven and twenty-five. A site was regarded as *indexed* if a search engine returned at least one page of the site. The last searches were performed in January 2008 at the end of the experimental part. In searching, the domain names were used as queries. Except for Voila, the *site command* of the search engines was used to restrict the search to the site in question. In all cases we ensured that the command worked as it was expected to work by comparing site searching to searching where phrases contained in sites were used as query keys. Voila does not allow site searching, so its results contained both relevant and irrelevant pages. Therefore, the results were reviewed manually to see if at least one page of the site in question was among the search results.

Google and Live Search are well-known major US-based search engines. A common view is that Google is the largest search engine. *Virgilio* is an Italian search engine. Because of its broad European coverage, it can hardly be regarded as a country-specific search engine, rather it is a regional search engine. The Finnish www.fi and French *Voila* are country-specific search engines. Their focus is on native sites, although they both index foreign Web pages.

The results were evaluated using the measure of coverage rate, which is defined as the proportion of sites returned by a search engine to all domain names in a test situation. For example, in the test situation, the French July sample divided by the results of the Google week eleven coverage rate was: ninety French sites returned by Google divided by 200 French domain names = 45.0% (see Table 7).

Altogether, 13,200 queries were entered into the search engines during the study (for each country 2x200 domain names and five search engines, which were queried at four different time points and 1200 Google searches at weeks 3 and 7).

Findings

Table 3 (the first experiment, May registrations) and Table 4 (the second experiment, July registrations) show the coverage rates of the test search engines over the three countries after eleven and twenty-five weeks following the registration days. As shown, in both experiments Google received the highest coverage rate among the five search engines. At week eleven, its coverage rates were 77.0% (the first experiment) and 66.7% (the second experiment). Between weeks eleven and twenty-five performance increased slightly: from 77.0% to 78.0% (Table 3) and from 66.7% to 77.0% (Table 4).

Perhaps surprisingly, *Virgilio* outperformed Live Search in both experiments. *Voila* received the lowest coverage rate in three out of four cases and *www.fi* in one case.

Tables 5 and 6 (the first experiment), Tables 7 and 8 (the second experiment) Tables 9 and 10 (summary tables containing the week eleven and week twenty-five results of both experiments) show the coverage rates of the five search engines by the home country of the sites. All six tables show that Google indexedU.S. sites more effectively than Finnish and French sites. In the first experiment, at week eleven its coverage rate for U.S. sites was very high: 98.0%. At week 25, the performance dropped to 91.0%. In the second experiment, the coverage rate increased from 85.0% to 88.0% between weeks eleven and twenty-five. A part of the decrease in the first experiment can be accounted for the removal of sites from the Web. The rest of the decrease cannot be attributed to the site removal, but the sites just disappeared from Google's index, or Google did not return them for some other reason. The same phenomenon was observed for the other search engines and for Finnish and French. This phenomenon has also been discussed by other researchers. Mettrop and Nieuwenhuysen (2001) observed that regularly submitted queries stopped retrieving documents that still existed on the Web. Bar-Ilan and Peritz (2004) found that search engines missed a remarkable number of previously located documents that still existed and that contained at least one of the search terms used in the queries.

Tables 5-10 show that Google's coverage rates were much lower for Finnish in particular for French. In the first experiment, Live Search performed better for Finnish than for U.S. However, the summary tables show that its coverage rate was highest for U.S. sites (Tables 9 and 10).

Live Search gave quite different results for the May and July sets, in particular in the case of U.S. sites/11 weeks: 18.0% for the May set and 50.5% for the July set. Also www.fi's performance at week 11 was inconsistent across the two sets. For the other search engines the results of the two sets were in fairly good agreement. The issue of search result inconsistency has been discussed in the literature. Thelwall (2008) discusses this issue and points out that search engines should not be viewed as mathematical “black boxes” that deliver logically correct results. Mettrop and Nieuwenhuysen (2001) examined the stability of search engine results based on their empirical findings they concluded that search engines miss documents in their result sets and are subject to changes in indexing policy. Intuitively, it seems that in particular the change of the search engine's indexing policy and the need to limit the burdening of the system may lead to substantial fluctuations in search results.

Virgilio received the second best coverage rate. It indexed more effectively Finnish and French sites than U.S. sites. Finnish www.fi indexed only Finnish sites and French *Voila* only French sites. This was expected, since they both focus on native Websites.

Search engine	Coverage % 11 weeks (n=600)	Coverage % 25 weeks (n=600)
Google	77.0	78.0
Live Search	20.3	41.5
Virgilio	41.0	47.7
www.fi	26.8	29.0
Voila	9.7	15.2

Table 3: Coverage rates of the test search engines. May registrations.

Search engine	Coverage % 11 weeks (n=600)	Coverage % 25 weeks (n=600)
Google	66.7	77.0

Live Search	36.0	45.8
Virgilio	41.5	52.8
www.fi	11.7	20.8
Voila	14.8	17.7

Table 4: Coverage rates of the test search engines. July registrations.

Search engine	Coverage % Finnish (n=200)	Coverage % French (n=200)	Coverage % U.S. (n=200)
Google	78.0	55.0	98.0
Live Search	24.0	19.0	18.0
Virgilio	63.5	37.5	22.0
www.fi	80.5	0.0	0.0
Voila	0.0	29.0	0.0

Table 5: Coverage rates of the test search engines by the home country of the sites. May registrations, 11 weeks.

Search engine	Coverage % Finnish (n=200)	Coverage % French (n=200)	Coverage % U.S. (n=200)
Google	81.5	61.5	91.0
Live Search	49.5	36.0	39.0
Virgilio	70.0	46.0	27.0
www.fi	87.0	0.0	0.0
Voila	0.0	45.5	0.0

Table 6: Coverage rates of the test search engines by the home country of the sites. May registrations, 25 weeks.

Search engine	Coverage % Finnish (n=200)	Coverage % French (n=200)	Coverage % U.S. (n=200)
Google	70.0	45.0	85.0
Live Search	27.5	30.0	50.5
Virgilio	57.0	35.5	32.0
www.fi	35.0	0.0	0.0
Voila	0.0	44.5	0.0

Table 7: Coverage rates of the test search engines by the home country of the sites. July registrations, 11 weeks.

Search engine	Coverage % Finnish (n=200)	Coverage % French (n=200)	Coverage % U.S. (n=200)
Google	83.0	60.0	88.0
Live Search	37.5	36.5	63.5
Virgilio	75.5	48.0	35.0
www.fi	62.5	0.0	0.0
Voila	0.0	53.0	0.0

Table 8: Coverage rates of the test search engines by the home country of the sites. July registrations, 25 weeks.

Search engine	Coverage % Finnish (n=200)	Coverage % French (n=200)	Coverage % U.S. (n=200)
Google	74.0	50.0	91.5
Live Search	25.8	24.5	34.3
Virgilio	60.3	36.5	27.0
www.fi	57.8	0.0	0.0
Voila	0.0	36.8	0.0

Table 9: Coverage rates of the test search engines by the home country of the sites. May and July registrations, 11 weeks.

Search engine	Coverage % Finnish (n=200)	Coverage % French (n=200)	Coverage % U.S. (n=200)
Google	82.3	60.8	89.5
Live Search	43.5	36.3	51.3
Virgilio	72.8	47.0	31.0
www.fi	74.8	0.0	0.0
Voila	0.0	49.3	0.0

Table 10: Coverage rates of the test search engines by the home country of the sites. May and July registrations, 25 weeks.

Figure 1 shows Google's coverage rate change at weeks three, seven, eleven and twenty-five for the Finnish, French and U.S. sites. As shown, for the U.S. coverage rate is high (97.5%) and near the maximum value even three weeks after the registration day. Coverage rates for the Finnish and French sites are remarkably lower at week three. For French sites, the performance increases almost linearly, while for Finnish there is a gap between weeks three and seven. Finnish sites seem to achieve a plateau at the end of the test period. The Finnish sites that disappeared from the Web after week eleven have a decreasing effect on the coverage rate at the end of the test period. It is likely that, similarly to U.S. results, Finnish and French coverage rates start to decrease, but later than U.S. sites (i.e., after 25 weeks)..

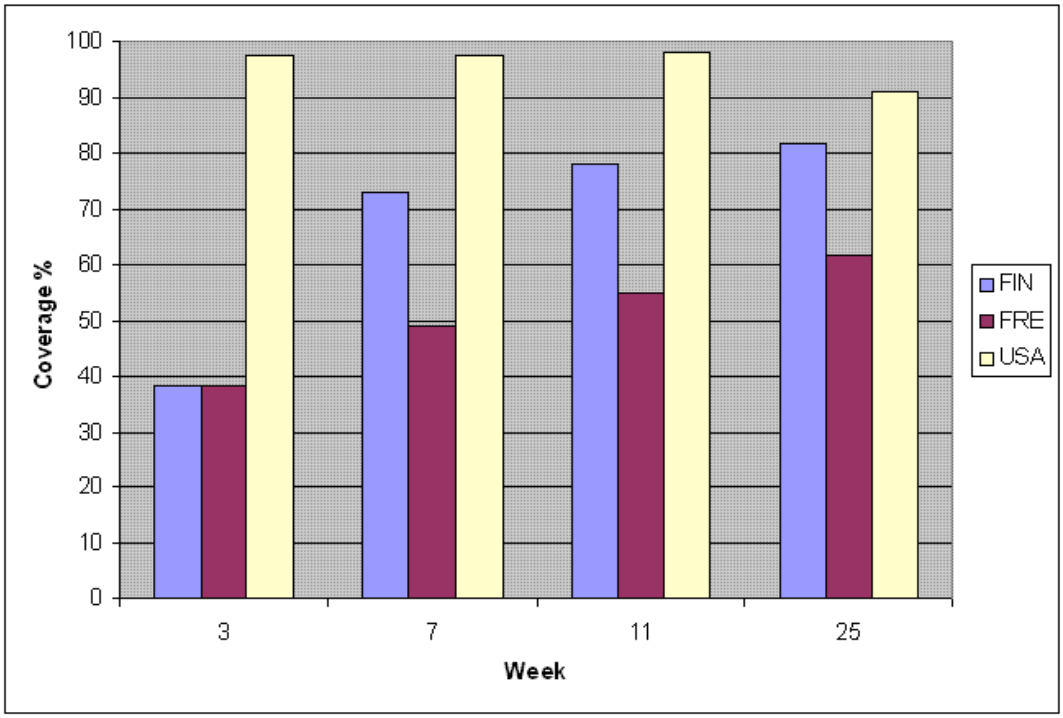

Figure 1: Google's coverage rate change for the Finnish, French U.S. sites.

Discussion and Conclusions

In this study, we investigated the effectiveness of five different search engines to index the Websites of new Finnish, French and U.S. domain names. The results showed that Google provided the highest coverage of the five search engines that US-based search engines Google and Live Search indexedU.S. sites more effectively than Finnish and French sites. The results are well in line with the earlier research findings, which showed that Google achieved the best coverage of several major search engines and that its coverage was better for U.S. sites than for sites from China, Singapore and Taiwan (Vaughan and Thelwall 2004 , Vaughan and Zhang 2007).

Our results also showed that *Virgilio*, www.fi and *Voila* indexed Finnish and/or French more effectively than U.S. sites. Apparently, there seems to be no problem here: US-based search engines index U.S. sites effectively, whereas European search engines focus on European sites. However, this is a matter of great concern due to the international nature of Google and Live Search. The European search engines only have local or regional significance. Google presumably is the largest search engine. A recent study showed that Google is the most popular search engine in the U.S., then followed by Yahoo! and Live Search (Sullivan 2006). Google's share of the search engine use was almost 50%. All these three engines are not only popular in the U.S., but almost all over the world. Besides the main engines, local versions, such as www.google.fi, are used widely. In many countries in Europe there is a great concern about the hegemony of US-based search engines, which indirectly decreases Europe's competitiveness. The concern has led to, for example, a European programme to develop multimedia, multilingual indexing and retrieval tools (Quaero). However, so far neither Quaero nor any other initiative has produced a serious rival for US-based search engines.

The issue of finding new information from the Web is as such an interesting and important issue. However, perhaps more important is the accumulation of the coverage differences generated in a day or a week. Naturally, during a long period of time a small daily difference in relative coverage results in a large difference when actual numbers are considered. Say, if a search engine indexes on average 500 sites a day more for one country than for another, in one month there will be, roughly, 15,000 sites more and in six months 80,000. The difference in the number of pages will be much higher, perhaps ten times or more.

What, then, are the reasons for the biased coverage of search engines observed in the study? We may only explain the results on a general level, because search engine companies are commercial organizations that seek to keep secret the detailed information on indexing and search processes. A search engine finds sites and documents that are not included in its database by following links from the documents that it already knows. Its *crawler* follows the links and downloads documents for the indexer, which constructs an index from the fetched documents. Words and URLs are extracted from the documents, the words are indexed and the URLs are added to the URL queue, which determines the order in which new documents are downloaded. Usually also other types of information, such as PageRank scores (Brin and Page 1998), are stored to make crawling and information retrieval more effective. PageRank rewards documents that have a large number of *inlinks* (i.e., links pointing to the documents) from documents that are themselves popular documents. So, crawling and the link structure of the Web play a central role in the inclusion of sites in the index of a search engine. However, they only play a role long after the registration of domain names, because the sites of new domain names generally do not have inlinks. For them, there are other inclusion mechanisms, discussed below. First we consider, however, an analysis that illustrates the role of inlinks in the search engine coverage of new sites.

We analysed whether Google's coverage of sites with inlinks was better than that of sites with no links pointing to the sites. In this analysis, the domain name data from twenty-five weeks after the July registrations were used and all three countries were considered. Google has a search option called *links* that searches for sites pointing to the URL used as a query, for example, the query *links: www.microsoft.com* finds sites pointing to Microsoft's Website. The sites of the test domain names were searched for using the links option for each country the following figures were calculated: (A) the percentage of test sites with at least one inlink and covered by Google among all test sites covered by Google, and (B) the percentage of test sites with at least one inlink and not covered by Google among all test sites not covered by Google. The results were as follows: U.S.: (A) 11.9% (176 sites covered by Google of which 21 sites were inlink sites) (B) 4.2% (24 sites not covered by Google of which one was inlink site). For Finnish and French the corresponding figures were: Finnish: (A) 22.9% (38/166) (B) 5.9% (2/34); French: (A) 26.7% (32/120) (B) 5.0% (4/80). As the figures show, the sites covered by Google have more inlinks than the sites not covered by Google. Clearly, inlinks contribute to the inclusion of sites in Google. It can also be seen that there are fewer inlink sites in the U.S. data than in the Finnish and French data. This finding suggests that the Web's link structure does not help to understand why Google was biased towards U.S. sites.

Biased coverage towards one country is caused, for example, by the fact that a search engine has learned or chosen to use new domain name data of that country (for example, such lists as we used to select the

test domain names) and has not learnt, or has chosen not, to use new domain name data of some other countries. This also means that the coverage of certain types of *new* sites by some search engine may change quickly if the search engine changes its indexing policy and uses additional sources on new domain names. However, the bias cumulated in a search engine because new sites from different countries have not been indexed equally in the past is more persistent. If a search engine with an imbalanced coverage changes its indexing policy towards a more balanced coverage, it would take a long time to obtain the balance.

In many cases the search engine's indexing policy simply favours sites of one country and sources containing information on them are scanned effectively. Actually, this is the core idea of country-specific search engines. In contrast to this, biased coverage in a global search engine is a matter of concern. It is a matter that decision-makers and Internet users should be aware of.

Many search engine companies allow manual submission of sites to the engine by the Webmasters of servers, who may submit a set of sites at one time. Different Web hosts can follow different submission policies. This may in part account for the obtained results. Also the publisher of a Website may promote the inclusion of the site in the index of a search engine. Moreover, there are companies that specialise in search engine submission and optimisation (the latter term refers to promoting the ranking of sites in search results).

Acknowledgements

This study was funded by the Academy of Finland (research projects 119600 and 125679).

About the author

Dr. Ari Pirkola received his PhD in 1999 in Information Studies at the University of Tampere, Finland. Since then, he has served as a researcher and teacher in the Department of Information studies at the University of Tampere. Currently he is working as a Finnish Academy research fellow. His research areas are information retrieval, in particular cross-language and multilingual information retrieval, language technology applications in retrieval, Web crawling and retrieval, and genomics retrieval.. He is a reviewer of several international journals and conferences and a board member of the National Language Technology Graduate School and the journal *Informaatiotutkimus*.

References

- Bar-Ilan, J. & Peritz, B.C. (2004). Evolution, continuity disappearance of documents on a specific topic on the Web: a longitudinal study of 'informetrics'. *Journal of the American Society for Information Science and Technology*, **55**(11), 980-990
- Barfourosh, A.A., Nezhad, H.R.M., Anderson, M.L. & Perlis, D. (2002). *Information retrieval on the World Wide Web and active logic: a survey and problem definition.* College Park, MD: University of Maryland, Computer Science Department. (Technical Report, CS-TR-4291) Retrieved 18 April, 2009 from http://www.lib.umd.edu/drum/bitstream/1903/1153/1/CS-TR-4291.pdf (Archived by WebCite® at http://www.webcitation.org/5g7zMQbLP)
- 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 Retrieved 18 April, 2009 from http://infolab.stanford.edu/~backrub/google.html (Archived by WebCite® at http://www.webcitation.org/5g804zaBw)
- Castillo, C. (2004). *Effective Web crawling*. Unpublished doctoral dissertation. University of Chile, Santiago, Chile. Retrieved 15 April, 2009 from http://www.chato.cl/papers/crawling_thesis/effective_web_crawling.pdf (Archived by WebCite® at http://www.webcitation.org/5g80dlmBZ)
- Gulli, A. & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. In *International World Wide Web Conference, Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan*, (pp. 902–903) New

York, NY: ACM Press. Retrieved 18 April, 2009 from http://www.di.unipi.it/~gulli/papers/f692_gulli_signorini.pdf (Archived by WebCite® at http://www.webcitation.org/5g80w3g1B)

- Introna, L. D. & Nissenbaum, H. (2000). Shaping the Web: why the politics of search engines matters. *The Information Society*, **16**(3), 169-186
- Lawrence, S. & Giles, L. (1999). Accessibility of information on the Web. *Nature*, **400** (6740), 107-109
- Lewandowski, D., Wahlig, H. & Meyer-Bautor, G. (2006). The freshness of Web search engine databases. *Journal of Information Science*, **32**(2), 133-150
- Lewandowski, D. (2004). Date-restricted queries in web search engines. *Online Information Review*, **28**(6), 420-428
- Mettrop, W. & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation*, **57**(5), 623-651
- Mowshowitz, A. & Kawaguchi, A. (2005). Measuring search engine bias. *Information Processing & Management*, **41**(5), 1193-1205
- Ntoulas, A., Cho, J., Cho, H.K., Cho, H. & Cho. Y.J. (2005). A study on the evolution of the Web. In *Proceedings of the US - Korea Conference on Science, Technology, & Entrepreneurship, University of California, Irvine, USA*. Retrieved 18 April, 2009 from http://oak.cs.ucla.edu/~cho/papers/ntoulas-evolution.pdf (Archived by WebCite® at http://www.webcitation.org/5g81UvqLS)
- 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
- Sullivan, D. (2006). Nielsen NetRatings search engine ratings. *Search Engine Watch*. Retrieved 15 April, 2009 from http://searchenginewatch.com/reports/article.php/2156451 (Archived by WebCite® at http://www.webcitation.org/5g849oFmq)
- Thelwall, M. (2008). 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. (2000). Commercial Websites: lost in cyberspace? *Internet Research: Electronic Networking Applications and Policy*, **10**(2), 150-159
- Van Couvering, E. (2004). *New media? The political economy of Internet search engines*. Paper presented to the Communication Technology Policy Section. *Conference of the International Association of Media & Communications Researchers (IAMCR)*, Porto Alegre, Brazil, July 25-30. Retrieved 18 April, 2009 from http://personal.lse.ac.uk/vancouve/iamcr-ctp_searchenginepoliticaleconomy_evc_2004-07-14.pdf (Archived by WebCite® at http://www.webcitation.org/5g84PUIvY)
- 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), article 7. Retrieved 18 April, 2009 from http://jcmc.indiana.edu/vol12/issue3/vaughan.html (Archived by WebCite® at http://www.webcitation.org/5g84cLoox)
- Vaughan, L. & Thelwall, M. (2004) Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, **40**(4), 693-707

How to cite this paper

Pirkola, A. (2009). "The effectiveness of Web search engines to index new sites from different countries" *Information Research*, 14(2) paper 396. [Available from 20th May, 2009 at http://InformationR.net/ir/14-2/paper396.html]