JIS Accepted for Publication
By the Journal of Information Science: http://jis.sagepub.co.uk

A three-year study on the freshness of Web search engine databases

Dirk Lewandowski[1]

Hamburg University of Applied Sciences, Hamburg, Germany

Abstract

This paper deals with one aspect of the index quality of search engines: index freshness. The purpose is to analyse the update strategies of the major Web search engines Google, Yahoo, and MSN/Live.com. We conducted a test of the updates of 40 daily updated pages and 30 irregularly updated pages, respectively. We used data from a time span of six weeks in the years 2005, 2006, and 2007. We found that the best search engine in terms of up-to-dateness changes over the years and that none of the engines has an ideal solution for index freshness. Frequency distributions for the pages' ages are skewed, which means that search engines do differentiate between often- and seldom-updated pages. This is confirmed by the difference between the average ages of daily updated pages and our control group of pages. Indexing patterns are often irregular, and there seems to be no clear policy regarding when to revisit Web pages. A major problem identified in our research is the delay in making crawled pages available for searching, which differs from one engine to another.

Keywords: search engines; online information retrieval; World Wide Web; index freshness

1. Introduction

Measuring the quality of Web search engines is a complex problem. While the focus is mainly on the retrieval effectiveness of the engines, we developed a general framework on search engine quality [1] that covers four areas:

- Index Quality: This points out the importance of the search engines' databases for retrieving relevant and comprehensive results. Measures applied include Web coverage (e.g., [2]), country bias [3, 4], and up to dateness [5].
- Quality of the results: This is the part where derivates of classic retrieval tests are applied. But, it should be asked which measures should be applied and if new measures are needed to satisfy the unique character of the search engines and their users.
- Quality of search features: A good set of search features (such as advanced search), and a sophisticated query language should be offered and work reliably (e.g., [6]).

[1] *Correspondence to: Dirk Lewandowski, Hamburg University of Applied Sciences, Faculty Design, Media and Information, Department Information, Berliner Tor 5, D – 20099 Hamburg, Germany. E-Mail: dirk.lewandowski@haw-hamburg.de*

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

- Search engine usability: This gives a feedback of user behaviour and is evaluated by user surveys, laboratory tests, or transaction log analyses.

The present study solely deals with a part of the index quality section. We believe that index freshness is an important part of quality measurement. Search engines should provide up-to-date information. We hope that the results from this study will be useful for being part of our overall search engine quality analysis.

Up-to-dateness with search engines derives its importance from several factors. First, there is the sheer size of the Web (see e.g, [2, 7-9] and its ever-changing contents. New pages are built, old pages are deleted, and links are changed, all at a high rate. But because of the growth of the Web, the number of old pages that no longer change has also increased significantly. Search engines have to find ways to show to the user pages that meet his or her up-to-dateness criteria. In some cases, older pages may be helpful, but in the majority of cases, one would assume that a user prefers current ones.

A study by Ntoulas, Cho, and Olston [10] found that a large number of Web pages are changing on a regular basis. Estimating the results of the study extrapolated over the entire Web, the authors find that there are about 320 million new pages every week. About 20 percent of the Web pages of today will disappear within a year. About 50 percent of all contents will be changed within the same period. The link structure will change even faster: About 80 percent of all links will have changed or will be new within a year. Although the absolute values may be out of date now, the results show how important it is for the search engines to keep their databases up to date. Huge and fast changes in the Web's contents are also reported in [11-14].

Bar-Ilan [15] studies the reasons for differing search engine results pages. Among others, she lists the following reasons directly related to up-to-dateness:

- Some of the search engines have several query engines or databases
- The index is partitioned
- When the crawler refreshes its database, some of the previously visited pages may be unreachable due to communication or server failure
- Fluctuations may be due to changes in the indexing policy of the search engines or in the size of the databases.

These are important factors that could explain inconsistencies in the results, as we will report below.

Ke, Deng, Ng, and Lee [16] give a good overview of the problems for search engines resulting from Web dynamics. Crawling and indexing problems resulting from Web dynamics from a commercial search engine's point of view can be found in Risvik and Michelsen [17].

Arguably every search engine user has experienced 404 errors with search engines (i.e., the page found by the engine links to a page that is no longer available). Newer studies [18, 19] show that the number of these errors is relatively low (for the top 20 results, between 2.2 and 6.5 percent and for the top 10 results, between 2.0 and 8.9 percent, depending on search engine), but even so, they are a major nuisance [20], pp. 179-180 [21] that results from unsuitable index freshness.

But how should search engines keep their indices up to date? It is quite clear that no search engine is able to update its complete index on a daily basis. This has economic as well as technical reasons, which we will discuss further in the next section.

It seems to be agreed among practitioners that search engines should index all pages in their indices in a cycle of one month. In our previous research, we were able to show that this does not hold true for the major search engines Google and Yahoo, even for pages whose contents are updated on a daily basis [5].But the question remains: Should search engines stick to such an update schedule, or can older pages be kept unindexed for a longer time? We will try to find an answer to this question, too.

This paper is organised as follows: First, we will give a concise review on the literature on search engine freshness, then we will state our objectives, and after that we will describe our methods. The emphasis is on the results of the current study in comparison to our previous study [5]. In the final section, we will draw conclusions and show areas for further research.

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

JIS Accepted for Publication
By the Journal of Information Science: http://jis.sagepub.co.uk

2. Literature review

The importance of freshness to search engines is often described and emphasised by the search engine vendors themselves [17, 22, 23]. It is a threefold problem that comprises issues with results ranking, with Web-based research, and with index freshness.

Freshness as a ranking factor is described by Acharya et al. [22]. There are lots of possibilities to use freshness factors for ranking: e.g., document inception date, content updates/changes, link-based freshness criteria, and changes in anchor texts. All major search engines apply freshness data into their ranking algorithms. But regarding the growing number of out-of-date Web pages, it is also important to recognise pages that no longer get updated. From the link structure surrounding these pages, one can assume whether a page is current or decays [23].

Freshness in Web-based research can be seen as a factor in information quality [24]. It is important for the searcher to get information that is current. With out-of-date information, the searcher will in most cases come to wrong conclusions for her work. Freshness can be a critical factor when a user wants to find only current information. Because of problems with determining the actual update of a Web document, search engines have problems in answering such date-restricted queries [6]. These problems result, at least in part, from the inability of search engines to differentiate between an actual update of the documents' contents and the mere change of design elements or minor alterations such as the current date and time, which is shown on some Web pages. Ntoulas et al. [10] distinguish between two measurements to determine an update of a Web document. On one hand, there is the *frequency of change*, which search engines currently use to determine an update. On the other hand, there is the *degree of change*, which is not used by the search engines sufficiently. The study finds that since there are often only minor changes in content, the use of the frequency of change is not a good indicator to determine the degree of change. Of course there may be exceptions to this (e.g., pages providing weather information), but for general text-based information pages, this seems to be true.

Some index freshness problems result from the general architecture of the database underlying the search engine. When a search engine uses batch indexing, the crawler builds the index, and when it has finished, it starts again to build a completely new index [17]. Therefore, search engines using this method are not able to dynamically add new pages to their indices. Some current results can be added in the process of the results presentation (e.g. news results), but the overall possibilities are limited. By contrast, incremental indexing does not have this problem as new pages can be added continuously to the index as they are found.

But another major problem appears here. The search engines have to define indexing patterns for each page in the index. When should the page be recrawled? With the batch indexing approach, the crawling process for all pages is the same. When the index is built, the crawler starts again to crawl all known pages. With incremental indexing, the search engine has to decide when to crawl each page. It is without a doubt true that not every page should be crawled with the same frequency. News Web sites change their contents often and should be crawled accordingly, while other pages stay the same for years after their inception.

The process for determining the update frequency can ideally be described as visiting the page, looking at whether the page is updated or not, and adjusting the update frequency to the frequency of actual updates. Therefore, the refresh interval adjusts permanently [17](p. 296). But with problems in determining actual updates, it is problematic for search engines to find the right intervals. Our study will ask whether the engines are able to find the right intervals, and if not, what the reasons may be.

But adjusting the crawl frequency to the actual update frequency is not the only way to determine which pages to crawl more often than others. Using only this approach is also problematic because all pages are treated solely on their updates, but not on their importance. Therefore, search engines can use link popularity to determine which pages should be updated more regularly. With limited resources, search engines are usually not able to crawl all pages according to their update frequencies and therefore focus on pages visited more often. The popularity of a page is usually measured with its link popularity, but other approaches such as click popularity could be used, too.

Bar–Ilan [15] proposes several new retrieval measures dealing with up-to-dateness, such as the ratio of broken links, the ratio of newly added pages, and the ratio of pages that are not known by any other search engine as of

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

yet. These measures have in common that they do provide indicators of freshness but do not examine a search engine's index as a whole.

We are aware of only two larger studies that deal with the average age of pages in the search engines' indices. One is the series of studies by Greg Notess [25], and the other our own investigation, which we continue in this article.

In the latest instalment of his studies, Notess [25] uses six queries to analyse the freshness of eight different search engines (MSN, HotBot, Google, AlltheWeb, AltaVista, Gigablast, Teoma, and Wisenut). Unfortunately the author gives no detailed information on how the queries were selected. For each query, all URLs in the result list that meet the following criteria are analysed: First, they need to be updated daily. Second, they need to have the reported update information in their text. For every Web page, its age is put down. Results show the age of the newest page found, the age of the oldest page found, and a rough average per search engine. In the most recent test [25], the big search engines MSN, HotBot, Google, AlltheWeb, and AltaVista all have some pages in their databases that are current or one day old. The databases of the smaller engines Gigablast, Teoma, and Wisenut contain pages that are quite older—at least 40 days.

When looking for the oldest pages, results differ a lot more and range from 51 days (MSN and HotBot) to 599 days (AlltheWeb). This shows that a regular update cycle of 30 days, as usually assumed for all the engines (e.g., recently by a leading industry publication[2]), is not used. All tested search engines have older pages in their databases.

For all search engines, a rough average in freshness is calculated, which ranges from four weeks to seven months. The bigger ones obtain an average of about one month, except for AltaVista, whose index has an average of about three months or older.

Notess's studies have several shortcomings, which mainly lie in the insufficient disclosure of the methods. It is neither described how the queries are selected, nor how the rough averages were calculated. The methods used in the described study were used in several similar investigations from 2001 and 2002. Results show that search engines are performing better in indexing current pages, but they do not seem to be able to improve their intervals for a complete update. All engines have quite outdated pages in their indices.

In our own study from 2006 [5], we used a selection of 38 German language Web sites that are updated on a daily basis for the analysis of the update frequencies of the major Web search engines. The cache copies of the pages were checked every day within a time span of six weeks. The search engines investigated were Google, Yahoo, and MSN. Only sites that display their latest update date or updated date information were used because Yahoo doesn't display the date when the cache copy was taken.

The analysis is based on a total of 1558 results for every search engine. We measured how many of these records are no older than 1 or even 0 days. It was not possible to differentiate between these two values because the search engines were queried only once a day. If there had been a search engine that updated pages at a certain time of the day it would have been preferred to the others. Therefore, it was assumed that a page that was indexed yesterday or even today is up-to-date in the cache.

Google handed back most of the results with the value 1 (or 0). The total number of 1291 records shows that 82.86 percent of the Google results were no older than one day. MSN follows with 748 (48.01 percent). Yahoo contains 652 (41.85 percent) one or zero days old pages in its index.

Also, the arithmetic mean up-to-dateness of all Web pages was calculated. Again, Google handed back the best results with an average age of 3.1 days, closely followed by MSN with 3.5 days, and Yahoo, who lags behind with 9.8 days. The use of the median instead of the arithmetic mean presents a different picture in which the competitors are closer together: Google and MSN have a median of 1 while Yahoo has a median of 4 days.

Another important point is the age of the oldest pages in the indices. While Google as well as Yahoo had several pages in their indices that were not updated for quite a long time, only MSN was able to completely update its index within a time span of fewer than 20 days. Since the research only focussed on Web pages that are

[2] http://searchengineland.com/070724-072605.php

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

updated on a daily basis, this cannot be proved for the complete index. On the basis of the findings it can be conjectured that Google and Yahoo, which both have outdated pages in their indices, will perform even worse for pages that are not updated on a daily basis.

To summarise the findings, Google was the fastest search engine in terms of index freshness, because many of the sites were updated daily. In some cases, there were outliers that were not updated within the whole time of the research or that showed some noticeable breaks in their updating frequency. In contrast to that, MSN updated the index in a very clear frequency. Many of the sites were updated constantly. Taking a closer look at the results of Yahoo, it can be said that this engine had the worst update policy.

3. Research questions

Based on the findings of our study reported above, we first wanted to know how the search engines under investigation perform over time. This part of the study repeats the investigations from the first study. Therefore, the first set of research questions is:

1. Are the update frequencies stable over the years? If not, how much do they differ?
2. Did the search engines fix their problems with their update patterns?

In addition, we wanted to formulate research questions that could not be answered in our first study, but were identified as interesting questions for further research. These are:

3. Is the update frequency for pages not updated daily different from the daily updated pages?
4. On which criteria do the search engines base their update policies?

We also added a new research question:

5. What is the delay from the crawl date to the inclusion in the index? Is bandwidth carefully used?

4. Method

In this section, we shortly describe the methods used in this study. As the current study continues the research [5], we refer to this text for more details on the choice of Web sites investigated and data collection. While the original study used data collected in 2005, we again collected data in 2006 and 2007.

4.1. Choice of search engines

In our previous study, we used three search engines (Google, Yahoo, and MSN). There were some requirements for search engines used in the study, as follows:

- Importance of the search engine: We only wanted to test the major search engines to get a view on the state of the art of search engine technology.
- Supply of cached copies of the crawled pages: For the purpose of our study, we needed search engines that make accessible a copy of the crawled pages and/or a precise crawl date for each page.

Only the major search engines Google, Yahoo, and MSN were able to supply the desired data. We reviewed the search engine market again in 2006 and 2007, respectively. There are no major changes, so we again used the same three search engines. Another new major player, Ask.com, could unfortunately not be included in our study because this search engine only supplies crawl dates (and cached copies) for some pages, but certainly not for the majority, let alone the pages from our test set.

In our first study, we used MSN (Microsoft Network) as one of the search engines investigated. This engine changed its name to Live.com in 2006. The change did not affect the technology, only the name and interface. Therefore, we stuck to the MSN name (which still is the name of Microsoft's [search] portal).

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

4.2. Selection of the pages and data collection

We used the same 40 pages[3] as in our 2005 study. A list of the sample pages can be found in the original article [5]. However, we had to make some changes due to changes with those pages. The reasons we had to make changes are as follows:

- Some pages are not updated on a daily basis anymore.
- Some pages now supply an automatic time stamp that is not included in Yahoo's cached copy because the stamp is done using a java applet. Therefore, Yahoo is not able to supply the crawl date of these pages.
- Some pages do not exist anymore.

These changes only affect a few pages and should not bias the overall results.

For 2006 and 2007, we added a control group of 30 Web pages to our data set. These pages are not updated on a daily basis, but have other update frequencies: some weekly, some only two or three times a month, and some irregularly or never. We used this data set to see how the search engines adjust their update strategies to the updated patterns of the Web pages themselves and to see how old pages in the engines' indices can be. In our first study, we found that there were some pages that were not crawled by the search engines for a very long time.

As in our first study, data collection was done manually. The reasons for this approach were that automated approaches through APIs do not produce reliable results [26] and that some of the search engines do not accept automated queries. Our data collection model is built on realistic user behaviour, as real users typed in the queries and examined the pages. The data collection method limits our study to a relatively low number of pages. However, the careful selection of pages across topics produces reliable results. One limitation is that we only used German pages in our sample. Therefore, our results can only give evidence for German pages. It would be possible that a search engine employing geographically distributed crawlers for different languages and/or countries could achieve better results for certain countries.

For each year, we investigated a time-span of six weeks. Data was collected from 15 February to 28 March 2005, from 15 February to 28 March 2006, and from 15 March to 25 April 2007, respectively. We do not believe that the changed data collection period for 2007 does have a temporal effect on the results. We carefully verified that every page in the sample was indeed updated every day during data collection.

5. Results

In this section, we first present the timeline for the pages investigated in the period from 2005 to 2007. Then, we will compare the daily updated pages to the control group of 30 pages with different update frequencies. Last, we will describe problems with the indexing delay of the different search engines.

5.1. Results for the pages that are updated daily

In a first step, we measured how many pages were not older than one or even zero days in our complete data set, i.e., the up-to-dateness information for every single page on every single day. We were not able to differentiate between the values one day old and zero days old because we queried the search engines only once a day. If there were a search engine that updated pages at a certain time of the day, we would have preferred it to the others. Therefore, we assumed that a page that was indexed yesterday or even today is up to date in the cache.

The results are shown in Fig. 1. While Google provided by far the best results in 2005 with 82.86 percent of all cached copies not older than one day, the results are far worse for the following two years. In 2006, Yahoo performs best with 73.13 percent of all cached copies current. While it still comes first in 2007, the results are much inferior (49.76 percent). MSN had the smallest number of current pages in its index in 2006 and 2007, while Yahoo came last in the first year of investigation.

[3] For the 2005 data, we had to omit two pages from the analysis due to technical reasons.

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

These results are quite surprising. One would assume that the search engines have their update policies and stick to them. Therefore, we did not expect major changes in performance over the years. But from the results, we cannot get a clear picture. We expect the main reason for worse performance in terms of freshness to be increased index sizes. In the time-span of our studies, all search engines under investigation increased their index sizes significantly. And as the Web continues to grow exponentially, probably search engines are not able to cope with its increased size.

A noteworthy item with Google is that the ratio of pages updated daily by this engine decreased largely from 2005 to 2006. Looking at the frequency distribution (Fig. 2), we find that at least for 2007, Google has a large number of pages that are updated daily, but that only occur in the searchable index after a delay of two days. We will discuss this phenomenon further in the section on update patterns.

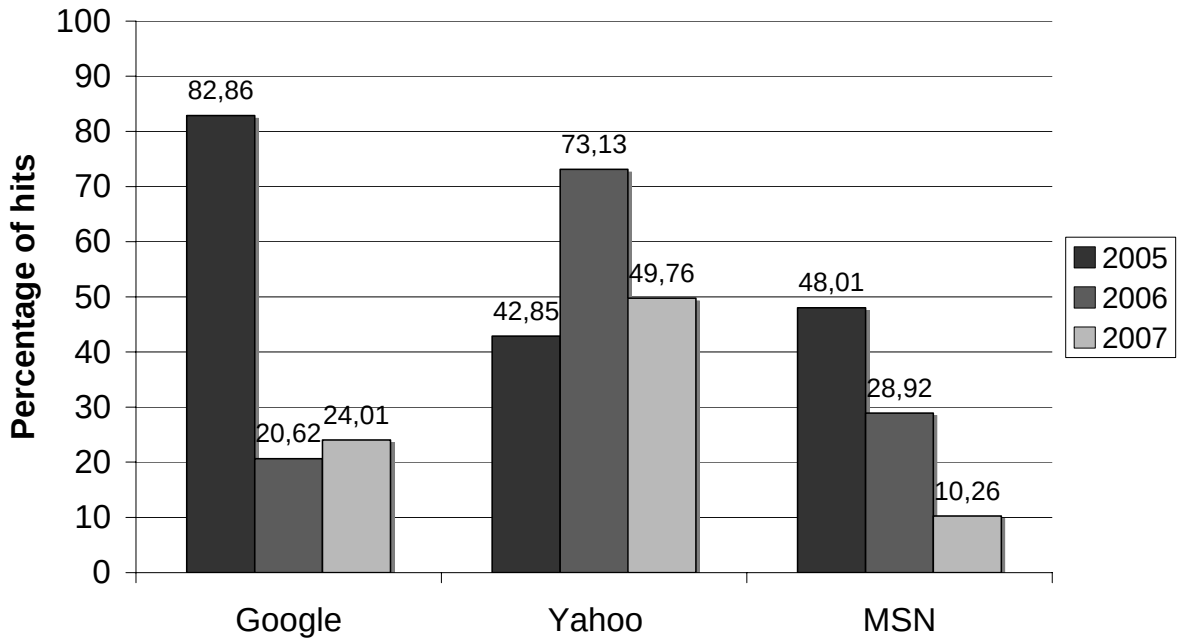

Fig. 1 ***Percent of pages that are up to date 2005–2007***

While the results presented so far only focus on the ratio of current pages in the search engines' indices, we also wanted to know how well the engines perform on average for all pages. Again, the results differ significantly from year to year (see Table 1). In 2005, Google performed best with a mean of 3.1 days. This search engine got worse the following year, and performed best in 2007 (mean: 2.2 days). A comparison between mean and median shows that in 2005 the mean resulted from some outlier in the data, while in 2006 and 2007, both values are similar.

Yahoo's performance changed over the years, too. In 2005, we found Yahoo performing by far the worst with a mean of 9.8 days and a median of 4 days. In 2006, the situation changed completely and Yahoo performed best of all engines with a mean of 2.3 and a median of 1 day. In the latest instalment, the data show a good performance of Yahoo with a mean of 2.7 and a median of 2.5. The data for all years show that while in the first two years there was a difference between mean and median, Yahoo has fewer outliers in 2007.

MSN's results got worse over the years, at least for the median age of the pages. Deviations between mean values for 2005 and 2006 are not significant, but in 2007, the performance was clearly worse.

Table 1 Average values for all pages that are updated on a daily basis

Search Engine	2005	2006	2007

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

Dirk Lewandowski

	Mean	Median	Mean	Median	Mean	Median
Google	3.1	1	5.6	5	2.2	2
Yahoo	9.8	4	2.3	1	2.7	2.5
MSN/Live.com	3.5	1	3.3	2	5.7	3

5.2. Newest and oldest pages found

In the next step, we look at the oldest and newest pages found in our data set of daily updated pages (results are summarised in Table 2). While all search engines are able to come up with at least some pages that are current (i.e., 0 or 1 day old), the results for the oldest pages found differ from engine to engine. Google gets better over the years (from 54 days old in 2005 to only 10 days old in 2007). Yahoo also performs better from year to year in our study, going from 62 days in 2005 to 26 days in 2007. MSN has nearly the same results for 2005 and 2006 with 17 and 16 days, respectively. Results get worse in 2007 with the oldest page found dated 30 days old. The results for the oldest pages found show that Google also clearly has the freshest index in terms of old pages.

Notess [25] reports that the oldest pages found in the latest instalment of his investigation were from 51 to 599 days old, depending on search engine. As the search engines used in these studies do not present the same search engine landscape as current, we cannot directly compare the results to ours. The only engine present in both studies is Google. Notess' 2003 study the oldest page found was 165 days old; this engine's results get better from year to year with a good result of 10 days in 2007.

Regarding our control group of pages with various update frequencies, we find that both Google and MSN have pages that were updated within our time-span of six weeks (and therefore are 0 or 1 day old). But when looking at the oldest pages found, we can assume that MSN is able to update its complete index within a time-span of 30 days, while Google has some outliers. The oldest page found in Google in 2006 is 253 days old, in 2007 175 days. This shows that Google is far from a current index of the Web.

Table 2 Oldest and newest pages found by the search engines

Search Engine	Newest Page Found		Oldest page found	
	Daily Updated Pages	Control Group	Daily Updated Pages	Control Group
2005				
Google	0	-	54	-
Yahoo	0	-	62	-
MSN	0	-	17	-
2006				
Google	0	0	29	253
Yahoo	0	-	32	-
MSN	0	0	16	19
2007				
Google	0	1	10	175
Yahoo	0	-	26	-
MSN	0	1	30	30

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

Dirk Lewandowski

5.3. Frequency distributions

It is interesting to see that in 2007, Google is able to update all daily updated pages within a time-span of 10 days. But as can be seen in Fig. 2, there is now a problem getting updated pages into the index fast enough. The peak of the curve is at two days. In 2005, Google was able to keep the majority of pages updated within a day, while in 2007 there is a delay of often two days from crawling to the availability of the page for querying the index. We will discuss this phenomenon in detail in Section 6.

The distribution for Yahoo (Fig. 3) shows clearly that while this engine has a large number of pages that are current or just a few days old, there is also a significant number of pages that are older, and some outliers that are very old.

The frequency distributions for MSN (Fig. 4) show a decreasing number of pages that are just a few days old, while the number of pages older than three days increases. However, the data for all years clearly show that MSN, as well as Google, updates all pages under investigation within a time-span of 20 days.

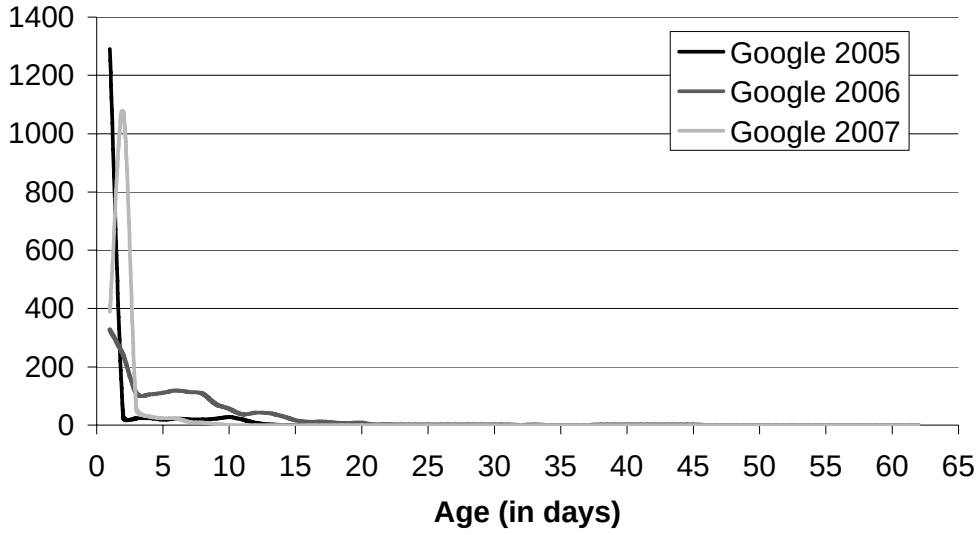

Fig. 2 Frequency of all Google values 2005–2007

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

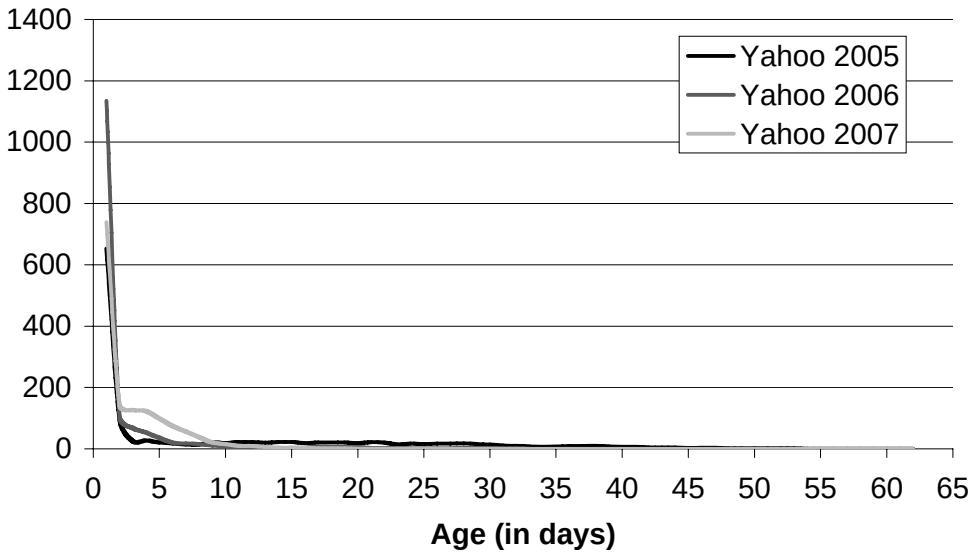

Fig. 3 Frequency of all Yahoo values 2005–2007

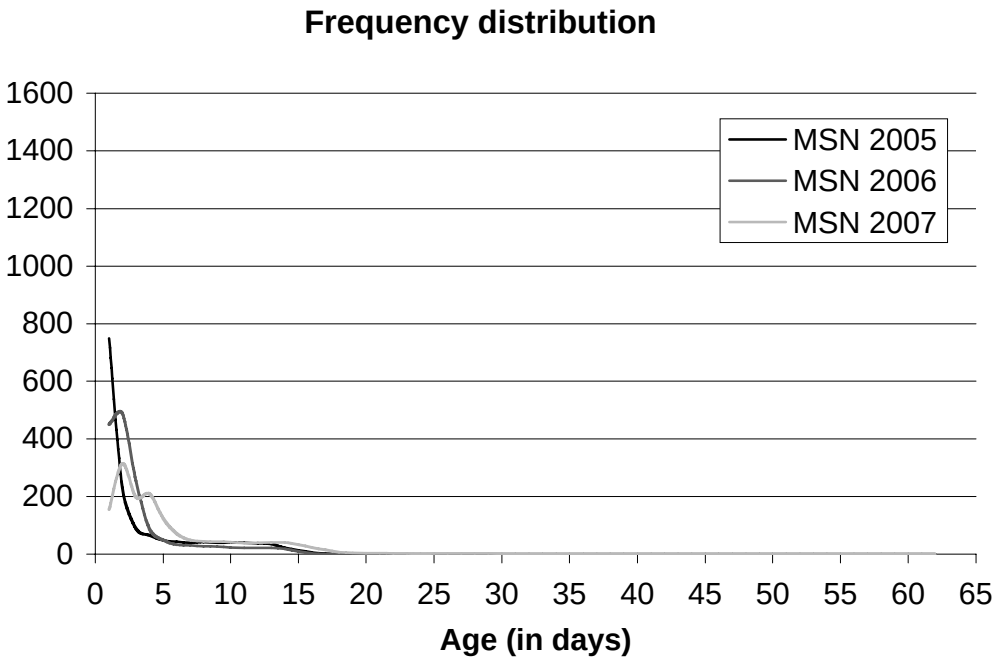

Fig. 4 Frequency of all MSN values 2005–2007

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

5.4 Results for the pages with a lower update frequency

For the control group of pages updated on a lower frequency, we were not able to obtain data from Yahoo. As noted earlier, this search engine does not provide a crawl date on the cache copy pages. As we did not know the actual update of the pages, we were not able to determine when Yahoo visited the pages, either. Therefore, our further analysis can only compare Google to MSN.

Table 3 clearly indicates that the update frequency for pages updated less frequently is different from the daily pages. As expected, search engines do have methods to ascertain which pages are updated more regularly than others. However, the mere fact does not imply how they do it.

The distribution for the irregularly updated pages is skewed for Google, as can be seen from the large differences of the mean and median for both years. This means that there are some pages in the control group that are crawled by Google more frequently, while others are only seldom crawled. Figs. 5 and 6 show that there was one page with a very high average value in 2006 (average: 232 days) and a few with higher values in 2007 (between 83 and 104 days).

For MSN, the situation is quite different. This search engine does not differentiate so much between the individual pages. Mean and median are comparable; i.e., most pages are updated with the same update frequency. This implies that MSN does take care to update its whole index within a certain time-span, while Google focuses on frequently updating pages considered as important, while deferring the update of others. This shows two different update strategies, one focussing on the whole index, one on a subset of important pages.

Table 3 Average values for all pages that are updated irregularly

Search Engine	2006		2007	
	Mean	Median	Mean	Median
Google	20.6	13.5	14.8	6
MSN/Live.com	7.7	8	9.3	9

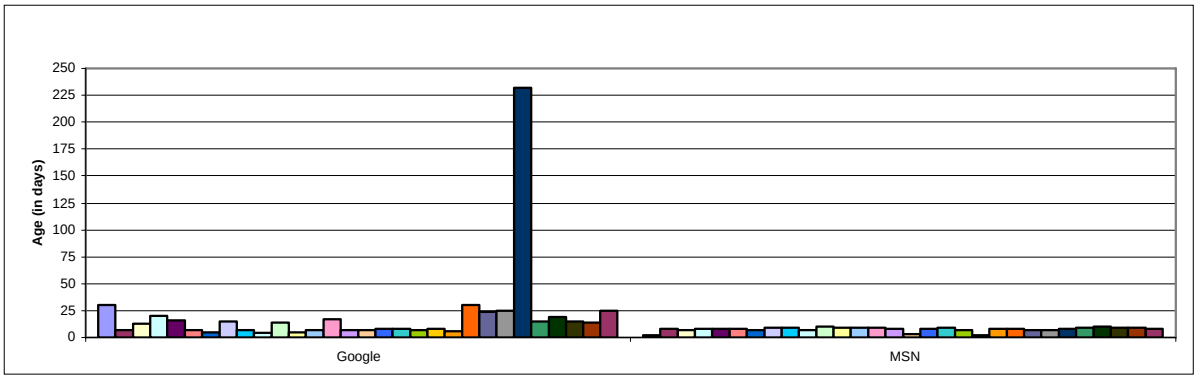

Fig. 5 Average values comparison showing individual control group pages 2006

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

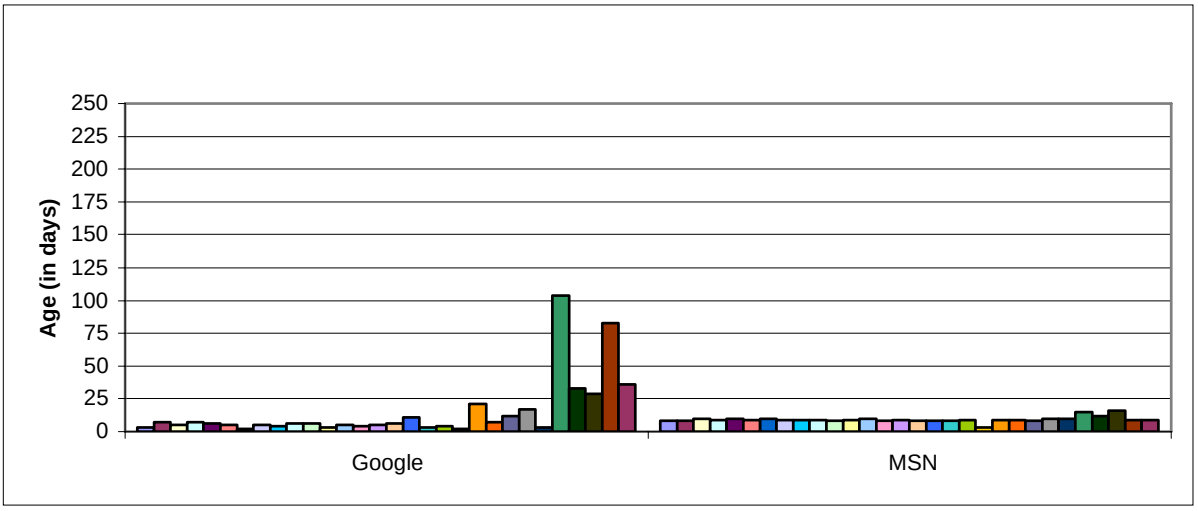

Fig. 6 Average values comparison showing individual control group pages 2007

The frequency distributions for the control group pages are shown in Figs. 7 and 8. It can be seen quite clearly that the update policies of Google and MSN differ greatly. While Google differentiates to a large degree between pages that should be updated more frequently and pages that do not seem important enough to be updated quite so often, MSN does not make this distinction to such a degree. Here, *all* pages are updated within a certain time-span. One could ask which approach is more useful to the user. With Google, it is more likely that the user gets the current version of the page, but for some pages, the cache copies are quite outdated. With MSN, cache copies are usually older, but a user will not be provided with a completely out-of-date version.

However, one must also consider that pages indicated as "not so important" are downgraded in the ranking. Therefore, the probability that a user stumbles upon a completely out-of date page is quite low. As this paper deals only with index freshness, we do not want to go further with this question.

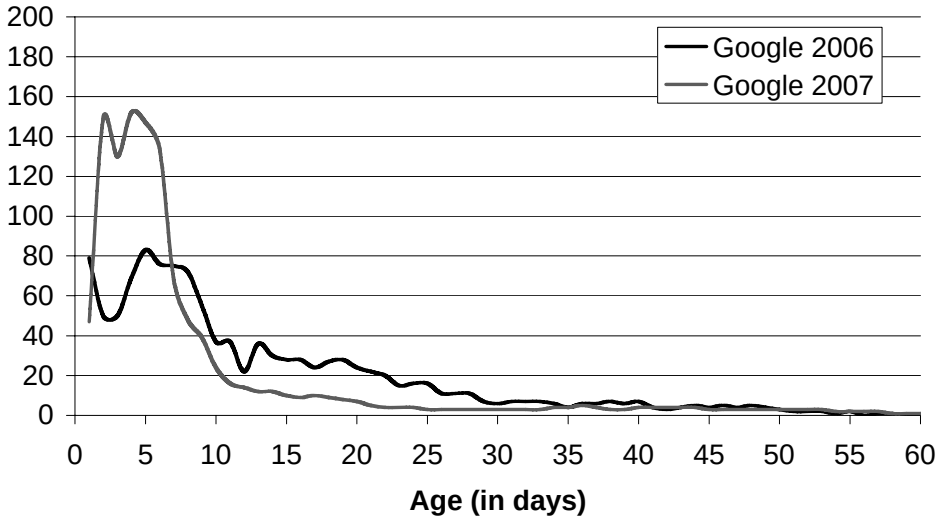

Fig. 7 Frequency for control group pages for Google

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

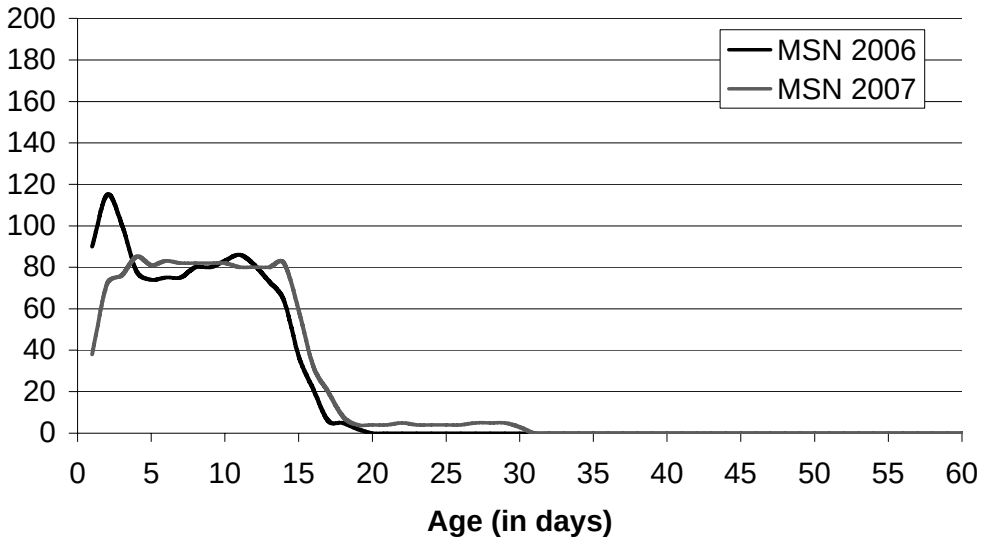

Fig. 8 Frequency for control group pages for MSN

6. Update patterns and indexing delay

As in our first investigation, we found irregular update patterns with all search engines. This holds true for the 2006 and the 2007 data, as well. As pattern features are similar for all years, we refer to the 2007 data in this section. We also focus on just two examples; a more general discussion can be found in Lewandowski et al. [5].

One would assume that the search engines have clear update cycles; e.g. daily, once a week, or once a month. But this is clearly not the case. Fig. 9 shows the update pattern for the German Wikipedia main page as an example. It can be seen that Google performs best but is not able to get the page into its index within less than two days in most cases, where Yahoo is able to get values of one day or less (at least in the last two weeks of the investigation). However, this search engine shows unreliable update patterns. In the first weeks of investigation, the cache copy does not get updated, while in the later weeks, it is recrawled quite often. Only MSN shows a regular update pattern. But even here, it stays unclear in which frequency the search engine wants to update the page. The first update is after 30 days, while the second is after just 16 days.

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

Dirk Lewandowski

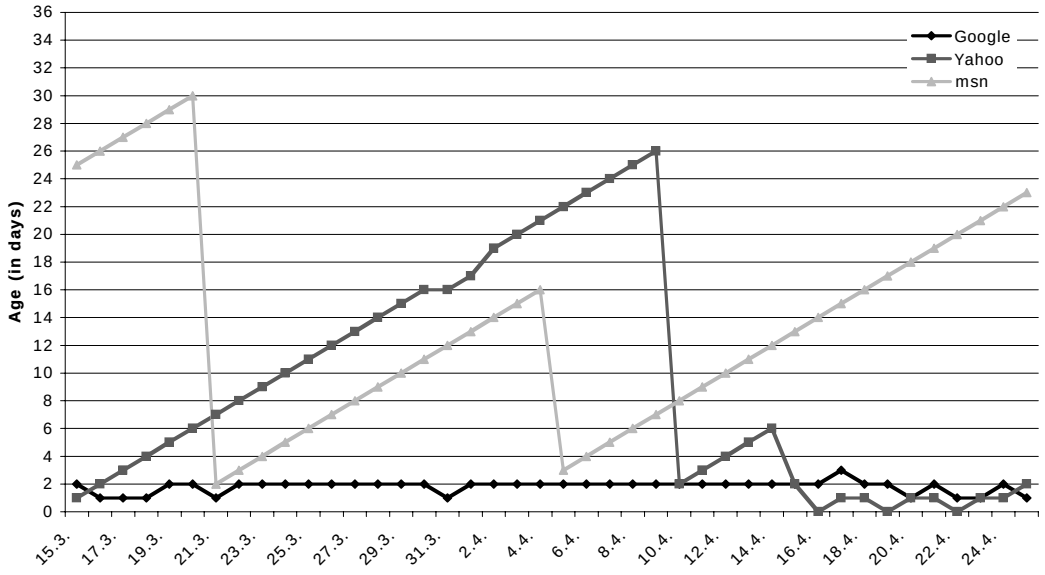

Fig. 9 Update pattern for German Wikipedia main page

As can also be seen from the example, Google provides the user with a new cache copy of the page every day. However, in most cases, the cache copy is two days old. This means that there is a certain delay between the crawling of the page and making it visible for the user (as reported in [17]). This also holds true for many other pages.

Fig. 10 shows that in 68 percent of cases, Google needs two days from crawling to making a page visible in search. With Yahoo, more than 50 percent of pages are available on the date they are crawled, while MSN needs one to two days.

The delays mean an enormous waste of bandwidth and thwarts up-to-dateness itself. When the engine needs two days to bring the crawled page into the searchable index, there is no need to crawl the page every day—just to get a two-day-old copy every day. To our understanding, it would only make sense to crawl a page on a daily basis if the engine is also able to bring the crawled pages into the index fast enough.

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

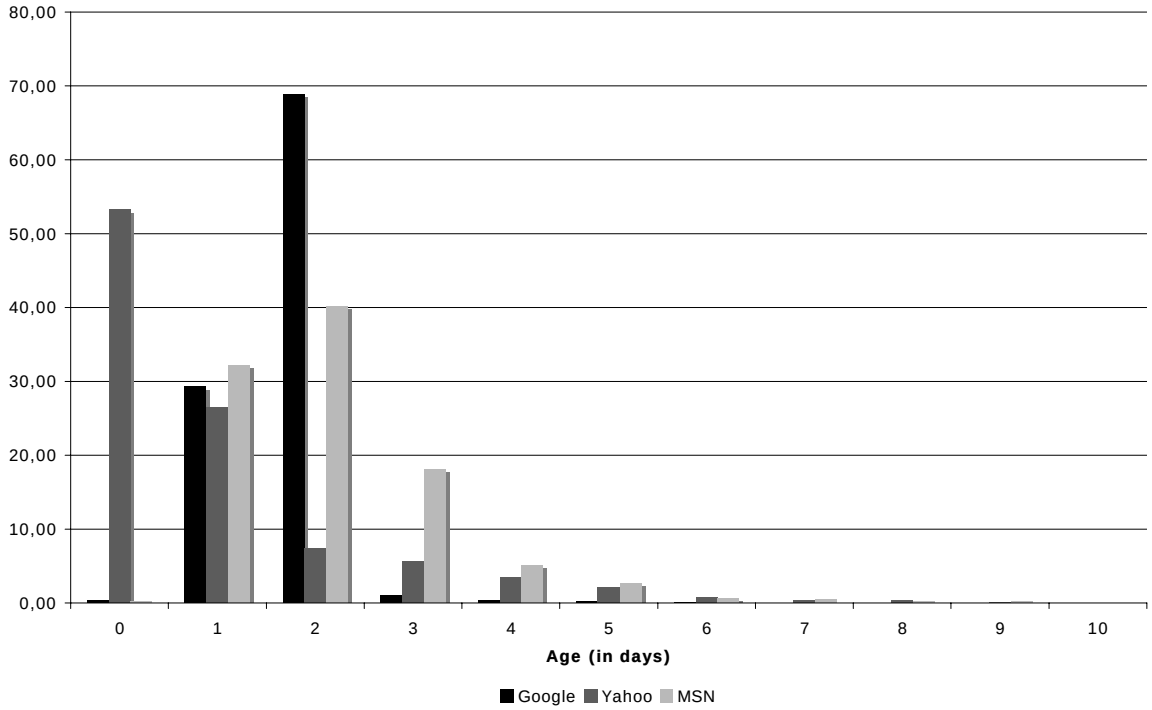

Fig. 10 Delay from crawling to the searchable index (percentage of pages)

7. Discussion

We will discuss our findings in regards to our research questions introduced in Section 3. Regarding our first research question (Q1: "Are the update frequencies stable over the years? If not, how much do they differ?"), we can say that we were not able to find constant update frequencies with any of the search engines under investigation. The frequencies with which pages are updated vary from one day to another, and therefore also from year to year. This makes it hard to give a clear recommendation for Web-based research. We are not able to recommend one search engine over the others, because one can never be sure whether the pages searched for are in the search engine's database in their most current version.

Regarding the update patterns (Q2: "Did the search engines fix their problems with their update patterns?"), we found that they stay rather unpredictable. However, we did not find pattern breaks, as discussed in our first study, to such an extent anymore. This shows that the engines fixed some of the problems that have to do with distributed indices.

Comparing the daily updated pages with the pages with a lower update frequency (Q3), our results confirm that search engines do distinguish between these (see also [17]). The update frequency for such pages is significantly lower than for the daily updated pages. However, there may be several other reasons why certain pages are updated more frequently; link popularity scores determined by the engines may be a prime reason.

The update policies of the different search engines (Q4) seem to be quite different. Google most significantly distinguishes between pages that should be updated very often and other pages that should not be updated for a longer time. MSN, on the contrary, does not show differentiation to such an extent, but updates its whole index within the time-span of a month. The Yahoo data does not allow for conclusions on the overall update policy because we were not able to obtain data for the irregularly updated pages.

Regarding the indexing delay (Q5), we found that this is indeed a problem for the search engines, mainly for Google. The delay is usually two days for this engine, while Yahoo is the only engine that is able to provide cache copies on the crawling date for a large number of pages.

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

8. Conclusions and further research

Our research shows that all search engines investigated have large shortcomings in updating their databases. None of the engines offers the ideal solution for the user (i.e., a comprehensive database of the Web that is updated according to the actual updates of the pages themselves). We found that none of the engines provides up-to-date copies even for the daily updated pages.

A question still remains to be answered: "How often should a search engine crawler revisit a certain Web page?" All engines do determine an update cycle from the actual update frequency of the page, combined with additional factors such as link popularity. While there are many works that recommend page visit policies of crawlers [27-31], the major search engines are not able to apply suitable policies or these do not scale to their index sizes.

However, update frequency is restricted by economic as well as technical factors. Still, search engines need to find a way to distinguish reliably between actual updates and merely small changes on the pages. An important goal is to assign the correct (content) update date to every page [6]. This will also solve the problem of ill-working date-restricted searches.

Regarding the pages that are not updated on a regular basis, search engines do waste effort on crawling the pages too often when unchanged content suggests not revisiting the page. On the other side, there are pages that are not crawled for a long time by a search engine (in the case of this investigation, Google). It still has to be determined which strategy proves better: Google's approach of not revisiting supposedly unchanged pages for a long time, or MSN's approach to revisit every single page in a certain time-span. Further research should focus on longer term observations of Web page changes and visits by the search engines.

References

[1] D. Lewandowski and N. Höchstötter, Web Searching: A Quality Measurement Perspective. In: A. Spink and M. Zimmer (eds.): *Web Searching: Multidisciplinary Perspectives*. (Springer, Dordrecht, 2008).

[2] A. Gulli and A. Signorini, The indexable Web is more than 11.5 billion pages. In: *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, Chiba, Japan (2005) 902-903.

[3] L. Vaughan and Y. Zhang, Equal representation by search engines? A comparison of websites across countries and domains, *Journal of Computer-Mediated Communication* 12(3) (2007) article 7.

[4] L. Vaughan and M. Thelwall, Search Engine Coverage Bias: Evidence and Possible Causes, *Information Processing & Management* 40(4) (2004) 693-707.

[5] D. Lewandowski, H. Wahlig and G. Meyer-Bautor, The Freshness of Web search engine databases. *Journal of Information Science* 32(2) (2006) 133-150.

[6] D. Lewandowski, Date-restricted queries in web search engines. *Online Information Review* 28(6) (2004) 420-427.

[7] A. Dobra and S.E. Fienberg, How Large Is the World Wide Web? In: M. Levene and A. Poulovassilis (eds.): *Web Dynamics - Adapting to Change in Content, Size, Topology and Use*, (Springer, Berlin, Heidelberg 2004) 23-44.

[8] P. Lyman, H.R. Varian, K. Swearingen, P. Charles, N. Good, L.L. Jordan and J. Pal, *How Much Information 2003? (2003).* Available at: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003 (accessed 26 November 2007).

[9] D. Sullivan, *Search Engine Sizes*. Available at: http://searchenginewatch.com/showPage.html?page=2156481 (accessed 26 November 2007).

[10] A. Ntoulas, J. Cho and C. Olston, What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In: *Proceedings of the Thirteenth WWW Conference, New York, USA* (2004).

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008

[11] W. Koehler, Web Page Change and Persistance - A Four-Year Longitudinal Study, *Journal of the American Society for Information Science and Technology* 53(2) (2002) 162-171.

[12] S.J. Kim and S.H. Lee: An Empirical Study on the Change of Web Pages. In: Y. Zhang, K. Tanaka, J.X. Yu, S. Wang and M. Li (eds.): *Web Technologies Research and Development - APWeb 2005: 7th Asia-Pacific Web Conference, Shanghai, China.* (Springer, Berlin, Heidelberg, 2005) 632-642.

[13] D. Fetterly, M. Manasse, M. Najork and J.L. Wiener, A large-scale study of the evolution of Web pages, *Software-Practice & Experience* 34(2) (2004) 213-237.

[14] M. Toyoda and M. Kitsuregawa, What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots. In: *Proceedings of the 15th international conference on World Wide Web*, (ACM Press, New York, 2006).

[15] J. Bar-Ilan, Search Engine Ability to Cope With the Changing Web. In: M. Levene and A. Poulovassilis (eds.): *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*, (Springer Verlag, Heidelberg, 2004) 195-215.

[16] Y. Ke, L. Deng, W. Ng and D.L. Lee, Web dynamics and their ramifications for the development of Web search engines, *Computer Networks* 50(10) (2006) 1430-1447.

[17] K.M. Risvik and R. Michelsen, Search engines and Web dynamics. *Computer Networks* 39(3) (2002) 289-302.

[18] J. Griesbaum, Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de, *Information Research* 9 (2004). Available at: http://informationr.net/ir/9-4/paper189.html (accessed 26 November 2007).

[19] J. Véronis, *A comparative study of six search engines (2006).* Available at: http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf (accessed 26 November 2007).

[20] M. Machill, C. Neuberger, W. Schweiger and W. Wirth, Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. In: M. Machill and C. Welp (eds.): *Wegweiser im Netz*, (Bertelsmann Stiftung, Gütersloh, 2003).

[21] N. Schmidt-Maenz and C. Bomhardt, Wie Suchen Onliner im Internet? *Science Factory/Absatzwirtschaft* (2) (2005) 5-8.

[22] A. Acharya, M. Cutts, J. Dean, P. Haahr, M. Henzinger, U. Hoelzle, S. Lawrence, K. Pfleger, O. Sercinoglu and S. Tong, *Information retrieval based on historical data (2005).* US patent number 20050071741.

[23] A.Z. Broder, Z. Bar-Yossef and S. Ravikumar, *Method and apparatus for assessing web page decay (2006).* US patent application number 10/995,770.

[24] S. Adams, Information Quality, Liability, and Corrections, *Online* 27(5) (2003) 16-23.

[25] G.R. Notess, *Search Engine Statistics: Freshness Showdown (2003).* Available at: http://www.searchengineshowdown.com/statistics/freshness.shtml (accessed 26 November 2007).

[26] P. Mayr and F. Tosques, *Google Web APIs - An instrument for webometric analyses? (2005).* Available at: http://eprints.rclis.org/archive/00003704 (accessed 26 November 2007).

[27] G. Pant and P. Srinivasan, Learning to crawl: Comparing classification schemes, *ACM Transactions on Information Systems* 23(4) (2005) 430-462.

[28] P. Srinivasan, F. Menczer and G. Pant, A general evaluation framework for topical crawlers. *Information Retrieval* 8(3) (2005) 417-447.

[29] M.D. Dikaiakos, L. Papageorgiou and A. Stassopoulou, An investigation of web crawler behavior: Characterization and metrics, *Computer Communications* 28(8) (2005) 880-897.

[30] V. Cothey, Web-crawling reliability, *Journal of the American Society for Information Science and Technology* 55(14) (2004) 1228-1238.

[31] J. Cho and H. Garcia-Molina, Effective page refresh policies for Web crawlers, *ACM Transactions on Database Systems* 28(4) (2003) 390-426.

JIS-0665 - v2 Received: 26th November 2007 Revised: 17th January 2008