JIS Accepted for Publication
Released: *6th October 2005 by Editor*
Accepted for publication in the Journal of Information Science: http://jis.sagepub.co.uk

The Freshness of Web search engines' databases

Dirk Lewandowski, Henry Wahlig and Gunnar Meyer-Bautor

Department of Information Science, Heinrich-Heine-University Düsseldorf, Germany

Correspondence to: Dirk Lewandowski, Heinrich-Heine-University Düsseldorf, Department of Information Science, Universitätsstraße 1, 40225 Düsseldorf, Germany. E-mail: dirk.lewandowski@uni-duesseldorf.de

Abstract.

This study measures the frequency in which search engines update their indices. Therefore, 38 websites that are updated on a daily basis were analysed within a time-span of six weeks. The analysed search engines were Google, Yahoo and MSN. We find that Google performs best overall with the most pages updated on a daily basis, but only MSN is able to update all pages within a time-span of less than 20 days. Both other engines have outliers that are quite older. In terms of indexing patterns, we find different approaches at the different engines: While MSN shows clear update patterns, Google shows some outliers and the update process of the Yahoo index seems to be quite chaotic. Implications are that the quality of different search engine indices varies and not only one engine should be used when searching for current content.

Keywords: search engines; Online Information Retrieval; World Wide Web; index quality; index freshness

JIS Editorial Process
Received: 19th Julv 2005
Revised: 6th September 2005
Version 2.0

1. Introduction

The numerous research papers dealing with the quality of Web search engines can be divided into two groups: The first deals with the quality of search engines' results, the second with the quality of the search engines' databases.

The quality of search results is usually measured with retrieval tests, e.g. [3, 5, 6, 9, 11, 12, 21, 24]. In more recent studies one finds that the retrieval effectiveness of the big search engines converges, but the overall precision is not that good.

In addition, some research focuses on the index quality. There are some studies dealing with the size of the search engines' databases [7, 8] which reveal that the search engines only index a small portion of the Web. Unfortunately, these papers are quite out of date, and further research should be carried out on this topic.

The depth of indexing is discussed in some older studies (cf. [4]) but there are no current results. Bias in Web crawling and therefore in the indices of search engines is discussed in some recent studies [2, 14, 23].

An important part of index quality lies in the freshness of the databases. Users looking for current information will find it only if the search engine's index is up to date. There are several search functions to evade the problem of out-dated indices, such as special news search engines [13] or blog search engines.

Our aim is to study the freshness of the databases of popular Web search engines. Users usually rely on their favourite search engine to provide them with the best results and give little thought about whether the desired information is in the index at all. On the basis of our findings we would like to give a recommendation to what search engine should be preferably used when searching for current information. First, we discuss the results from past studies dealing with date-issues. Then we present our study of the freshness of search engines' databases, which will be discussed in detail.

2. Related studies

2.1. Notess 2001-2003

Notess [15] 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 are analysed which meet the following criteria: First, they need to be updated daily. Second, they need the reported update information in their text. For every webpage, 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 [15], 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, 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 reach an average of about one month except AltaVista of which the index with an average of about three months is older.

Notess' study has 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. Since only a small number of queries were used, only 10 to 46 matches per search engine were analysed. This amount is simply too small to get more than just exploratory results. Another shortcoming is that the author does not explain how he determined the date when the different search engines indexed a page. For some engines the cache could be used, but not all of them offer this kind of function. Notess' research is discussed here simply because it has been – at least to our knowledge – the only attempt similar to our approach so far.

The methods used in the described study were used in several similar investigations from 2001 [19] and 2002 [16, 17, 18]. Results show that search engines are performing better in indexing current pages, but they do not seem to able to improve their intervals for a complete update. All engines have quite outdated pages in their index.

2.2. Lewandowski 2004

In a study testing the ability of search engines to determine the correct date of web documents, Lewandowski [10] finds that the major search engines all have problems with this. He uses 50 randomly selected queries from the German search engine Fireball, which are sent to the major search engines Google, Yahoo and Teoma. These engines were selected because of their index sizes and their popularity at the time of the investigation. All searches were done twice: once without any restrictions, once with a date-restriction for the last six months. For each query, 20 results were examined for date information. The study reveals that about 30-33 percent of the pages have explicit update information in their content. This information was used to compare the non-restricted with the date-restricted queries.

The number of documents from the top 20 list that were updated within the last six months was counted and was defined as the up-to-dateness rate. The proportion of these documents, out of all the documents, was defined as the up-to-dateness rate. The corresponding sets of documents retrieved by the simple search, as well as by the date-restricted search, were calculated. The up-to-dateness rates for the simple search are 37 percent for Teoma, 49 percent for Google, and 41 percent for Yahoo. For the date-restricted search, the rates are 37 percent for

Teoma (which means no improvement), 60 percent for Google, and 54 percent for Yahoo. Taking this into consideration even Google, which proved to be the best search engine in this test fails in 40 percent of all documents. All in all the study shows that all the tested search engines have massive problems in determining the actual update of the found documents. But this data could be very useful for the indexing and even the ranking process (cf. [1]).

The study recommends using information from several sources to identify the actual date of a document. The following factors should be combined: server date, date of the first time the document was indexed, metadata (if available), and update information provided in the contents of the page [10].

2.3. Ntoulas, Cho and Olston 2004

The problems described in Lewandowski [10] could 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, Cho and Olston [20] distinguish between two measurements to determine an update of a Web document. On the 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 the 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, such as pages providing weather information, but for general text-based information pages, this seems to be true.

Furthermore the study can prove that a large amount of Web pages is changing on a regular basis. Estimating the results of the study for the whole 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 be new within a year.

The results show how important it is for the search engines to keep their databases up to date.

2.4. Machill, Lewandowski, and Karzauninkat 2005

In a "reaction test" [13], this study investigates how fast the several engines index news content. The test was designed to analyse the speed of search engines in integrating events of topical interest into their news indices. Nine search engines were tested one hour, three hours and five hours after an event appeared in the Reuters news ticker. Just one query was used, "Hubschrauber Absturz im Irak" (helicopter crash in Iraq).

Results show that Google as well as AltaVista and Yahoo provided the best reaction time of less than an hour. All other search engines with the exception of T-Online (a German portal) were able to provide first articles after

three hours. Yahoo, AltaVista and Google were in the lead with regard to the number of hits as well as to the impact of the results.

A strong limitation of the test is the use of just one query. Apart from that it is obvious that a good reaction time is needed to keep the news indices current. It would be very difficult to integrate the news results into the regular index of crawled Web pages. Search engines therefore separate the two databases. Additional news indices would not be needed if the engines were able to integrate the current news into their regular databases.

3. Objectives of this study

The previous chapter proved that the testing of the freshness of search engines´ databases is still new in information science. As we have seen there are just a few studies published and they all focus mainly on other aspects in the broad topic of date information. The representative daily measuring of a search engines index' up-to-dateness, as planned in this research, has never been ran in such an test environment so far.

Keeping that in mind we are aware that our study has to carry out basic research especially in terms of its method. We attach particular importance to the experimental setup and description of the workflow, which tries to find the best way to test the up-to-dateness of the web page indices. Our effort is that our method can play the role of a model for similar following studies on the same topic.

In terms of content, this paper concentrates on more general questions and compile some basic facts about the update process of the search engine indices. With our tests we would like to find out in which frequency the search robots update their indices: Are there any specific intervals for every single webpage or – to put it more general – are there any clear intervals at all? So far, assumptions have speculated that the search engines update their whole index within a period of thirty days. These assumptions were proved false by Notess [15] but this gross theory has to be double checked in our research because of possible improvements since Notess' studies.

Certainly, the comparison of the three competitors is an important focus of our research. Which similarities and differences can be stated between the search engines and what do they tell us about the quality of the index? Can Google confirm its role as the perceived technology leader in search engine technology in terms of its up-to-dateness as well?

As already mentioned our work will define some basic statements and raise more specific questions for later research. It means that we can answer, *if* the search engines update their indices properly, but not *why* this works better or worse with some pages or engines. These questions have to be answered in later research, which concentrate on certain phenomena. For these points, our general data basis is simply not detailed enough.

4. Method

4.1. Selection of the pages

Our first goal was to find 40 German websites, which are updated on a daily basis. The number of 40 websites was chosen since it seemed to hand back the best relation between reliable results on the one hand and acceptable work efforts for us testers on the other hand. From this basic set, we had to omit two pages in the course of our study due to technical reasons.

The first restriction was that all websites had to show the date of their last update within their content. This restriction had to be made since not all of the search engines publish the information when the cache version was taken. While Google and MSN displayed the date above the record taken from the cache, Yahoo does not. To clarify the date of that page we had no other chance than to get the date from the webpage itself. To collect this information, we simply needed an automatic date generator, which displays the actual date somewhere in the content. Many news portals, for example, show the actual date and time on top of every of their pages. Another study [10] already proved that the search engines cannot distinguish between minor changes (e.g. change of date) and changes that affect the content. We were mainly interested in the content of the page and we wanted to get a representative image of daily-updated-websites for our research. Therefore we formed four groups with nine or ten websites in order to reflect the actual situation on the web: 18 news portals (group 1 and 2), 10 scientific oriented pages (group 3) and 10 special interest websites (group 4); although we were conscious of the fact that there are some of the overlap between the groups.

The next paragraphs present these groups. Details on the sites can be found in appendix A.

4.1.1 National news portals

The first group consists of (like the second one) nine clearly news-oriented websites. The difference between these two groups is that this first group concentrates on news portals with a supra-regional character and mainly on national and international news. The websites satisfy common information needs of visitors from completely different backgrounds and exceed the members of the second group in importance and traffic.

4.1.2 News portals with regional character

This second group with nine websites also focuses on news-oriented but more regional websites. The news services offered there may include national news, but mainly concentrate on regional or local information.

4.1.3 Scientific oriented websites

The third group consists of scientific oriented websites. Main focus of these pages is to transfer specific academic news or background information. Main addressee is not the public in general, but people who are interested in this topic. Traffic and importance decisively depend on the single topic.

4.1.4 Special interest websites

The fourth group contains 10 special interest websites. The topics range from specific ones on certain leisure time activities such as camping to more general topics as for example online flirting. The traffic and addressees on these websites also vary depending on the topic.

4.2. Selection of the Search Engines

After choosing the websites for our research we had to decide which search engines should be examined. We wanted to get a representative and comprehensive overview of the key players on the German market.

In order to fulfil these conditions, we first had to get an overview of the current situation on the German search engine market. When analysing the dependence of the different companies we found out that only Google.de, Yahoo.de and MSN.de run their own indices. All other big players in Germany such as T-Online or AOL get their results from one of these companies.

Empirical studies of the user behaviour towards German search engines are rare and mostly quite old. The latest publication on that topic from Machill et al. [12] dates back to the year 2003. We therefore observed current international studies, which are frequently updated on searchenginewatch.com. Nielsen Net Ratings [22] is one of them. It measures the search behaviour of more than a million representative users worldwide on a monthly basis. The results convey a clear message: Google, Yahoo and MSN together share more than 80 percent of the complete market. In March 2005, Google got a share of 47 percent, Yahoo 20.9 percent and MSN 13.6 percent. All other competitors were in a negligible position.

Our hypothesis was that this worldwide trend is also valid for Germany. In order to prove this assumption, we used a special empirical instrument on the web: WebHits (www.webhits.de), the leading German company offering web counters for private and commercial homepages. One of the services they offer is to determine which search engine brought the user to the websites using a WebHits counter. The variety of the customers using WebHits is very large and includes small private websites but highly frequented commercial pages as well. Generally, the updated statistical output is based on approx. 50.000 queries per day. The results show an even clearer, but finally comparable result: 80 percent use the search engine Google, whereas MSN and Yahoo share the second position with about 5 percent each.

These clear findings affirmed our hypothesis and encouraged us to run the study with the three big independent search engines on the German market: Google.de, Yahoo.de, MSN.de.

4.3. Workflow: How to measure freshness?

Our research tried to find out whether these theoretical statements can be sustained in reality. In order to get an answer, we had to find out how fast our 38 daily updated pages are updated in the Google, MSN and Yahoo index. We will call this value the "freshness rate".

The measuring of the index up-to-dateness seemed to be not that easy, since the age of the index versions is not directly displayed by the search engines. Although the result list is based on the internal index, the user is lead to the real website on the web when he or she clicks on one of the items. To avoid that forwarding, we developed the following workflow, which determines the "freshness rate", which is defined as the time lag between the internal cache version and the real version of a web page:

1. The complete URL has to be entered in the query box of the search engine.
2. In the displayed record, the link "Cache-Version" (Google), "im Cache" (Yahoo) or "zwischengespeicherte Seite" ("Cached page", MSN) had to be followed. This link refers not to the real webpage, but to the current version of the same page in the internal index.
3. The displayed page shows the cache version and some further information in a text above. MSN and Google include the date and time when the cache version was taken, whereas Yahoo does not show any date or time at all. That forced us, as already mentioned, just to select websites with a date record in the content itself. However, we finally got exact dates for every cached webpage. We first thought about using the http header of the cached pages for determining the actual date, but this did not work since the header always shows the current date.
4. In order to finally get the freshness rate, the dates of the cache versions must be set in relation to the actual date of that day. A cache version from March 8th for example has a freshness of two (days) on March 10th. The freshness rate of a search engine's database is the average freshness of all pages examined.

The described workflow led to a complete statistical output of 4788 specific freshness rates (38 webpages x 3 search engines x 42 days).

4.4. The outward setting

Besides these theoretical questions, we also had to define some further practical settings at the beginning of the test. The first one concerned the length of our research: We finally decided to run our tests for exactly six weeks or 42 days. This time seemed adequate enough to lead to reliable results on the one hand, with manageable proportions on the other. The study was eventually carried out from February 15th to March 28th 2005.

Moreover, we opted to repeat our tests on a daily basis – no other frequency would have guaranteed us consistent results. An important factor in this context was to lay down a fixed time at which the tests had to be repeated daily: Otherwise disparities in the results would have certainly emerged, since a cache version not being updated on 8 a.m. may have been already renewed at 8 p.m. the same day. This result would have led to two different freshness rates.

We laid down 6 p.m., which was not meant as a strict statue but more as a reference value. The time period for our test was finally extended from 5 p.m. to 8 p.m. This was no problem since our pretests clearly showed that all search engines tend to update their caches once a day – mostly at night time (Central European Time).

4.5. Pretest

An extensive pretest was run for five weeks beginning on December 18th and ending on January 19th. In this test we examined the same search engines Google, Yahoo and MSN, but limited the number of pages to the five websites web.de, sportschau.de, idw-online.de, rp-online.de, and uni-duesseldorf.de.

The systematic of the pretest later determined the workflow of the main test. It was primarily carried out by a group of four students whereas two of them took part in the main test as well.

The main purpose of the pretest was to get a first impression, whether it is worthwhile to carry out a more comprehensive study with a larger amount of websites. Besides, the daily tests were intended to gain first-hand experiences on the practicability of the workflow.

The results of the pretest produced a divided picture: Google and MSN reached constant freshness rates of about 1.0 and 2.3 days on average, whereas Yahoo caused several problems. On some days, the cache was completely unreachable, on other days the results varied extraordinarily. We were astonished by the continuing phenomenon that Yahoo showed on one day old cache versions and on the next day versions that were nine or ten days old. We concluded that at least Yahoo could not guarantee constant updates of its index. This phenomenon of sudden "refreshment gaps" also occurred at MSN, although not that extent. These curious results encouraged us to examine this subject on a broader scale in an additional main test.

Furthermore the pretest proved that the four-steps-workflow was the right instrument. Therefore it was applied to the main test as well. Technical problems were only caused by other factors, since esp. Yahoo removed the display of the cache for some sites during the test.

4.6. Main test

The main test started on February 14th and was, as already mentioned, conducted for 38 websites, 3 search engines and 42 days. After all, this meant that the final analysis would be based on 4674 single records (38 x 3 x 42).

Unfortunately this number of records had to be reduced because of some insurmountable technical problems we already recognized during our pretest: From time to time, the search engines (esp. Yahoo but never Google) removed the display of the cache from their website over night. For some sites like reuters.de, this phenomenon occurred nearly daily. One day the cache was displayed, the next it was not. This inaccessibility of the cache version reduced our total statistical output to 4572 records.

Another problem led to a further reduction since one of the sites changed its layout during the tests. The portal http://www.liebesalarm.de removed the date display from its website on March 10th, so we could not assess the age of the cached version at Yahoo anymore. This limited our statistical output to a final number of 4556 records.

Besides that, no further technical or other problems occurred during the test.

5. Results

The presentation of our results contains three parts. The first one will point out some general aspects which are valid for the complete research. We will then focus on the four groups, which were selected according to their topic and popularity (see section 4.1. ff.). We will hopefully prove the assumption that the update frequency has something to do with the content of the website. Part three will finally concentrate on the indexing patterns of the three search engines and observe the technique with which the search robots update their indices.

5.1. Overall results

First of all we wanted to find out which search engine provides the best results. When searching the Web every search engine usually hands back a large amount of results. But how do we know that we got the latest information on our topic? Perhaps one page was updated just yesterday with some really new aspects. An ideal search engine would update its index on a daily basis. In our first analysis we want to examine how close each search engine gets to that ideal.

In our research we got an overall data set of 1558 results for every search engine. In this evaluation, we measure how many of these records are not older than 1 or even 0 day. We were not able to differentiate between these two values because we queried the search engines only once a day. If there was a search engine that updated pages at a certain time of the day we would have preferred it to the others. Therefore, we assume that a page that was indexed yesterday or even today is up-to-date in the cache.

As shown in figure 1, Google hands back most of the results with the value 1 (incl. 0). The total number of 1291 records shows that 82.86 percent of the Google results were not older than one day. These results are much better than those of the competitors and are an obvious sign for the premium quality of the Google index.

MSN follows with 748 (48.01 percent). Yahoo contains 652 (41.85 percent) one or zero days old pages in its index.

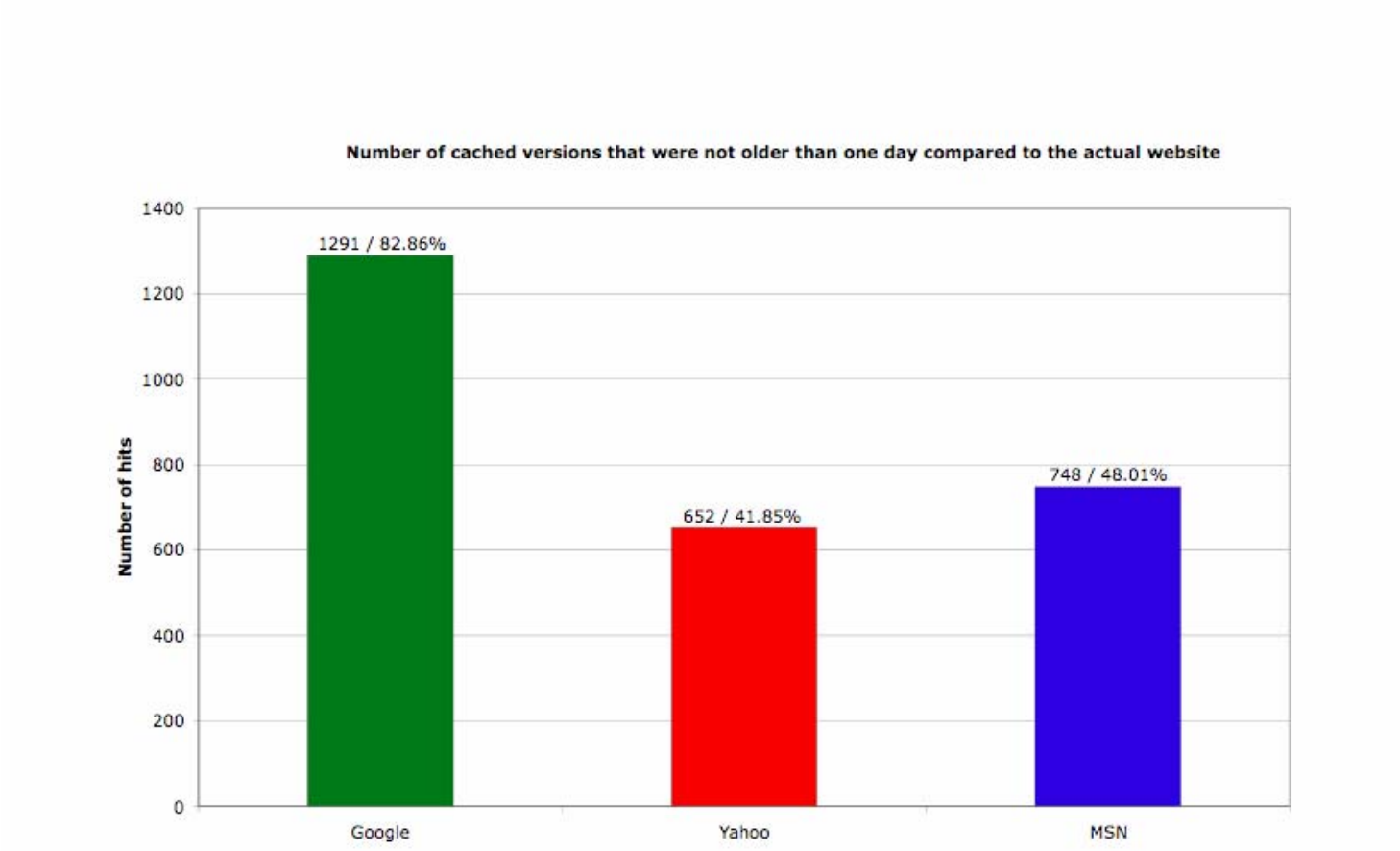

Fig. 1. Number of cached versions that were not older than one day compared to the actual website.

Obviously, the percentage of current websites cannot be the only indicator to measure the overall up-to-dateness of a search engines index. Moreover, we have to consider the older results as well to get a complete overview.

Therefore we calculated the arithmetic mean up-to-dateness of all web pages in our research. Again, Google hands back the best results with an average age of 3.1 days, closely followed by MSN with 3.5 days and Yahoo is way behind with 9.8 days. The use of the median instead of the arithmetic mean draws 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.

How is it possible that both measurements hand back different results? Considering only the arithmetic mean Google seems to be the absolute leader whereas the median reveals that there hardly is a difference between Google and MSN.

Figure 2 provides the answer to this question: It shows the arithmetic mean of all 38 web pages for all three search engines at one glance. For most of the pages the freshness of the Google results is quite acceptable but there are a few pages which reach an average of six days and two pages with an average between 25 and 27 days.

Due to these outliers the overall average of Google falls back to just 3.1 days and therefore endanger the overall top position of the search engine, since in contrast to Google MSN in does not show such outliers. The number of current websites may be less here, but the number of outliers as well. Nevertheless Google keeps its position as the search engine with the freshest results.

The third competitor Yahoo cannot reach the position of the two others. The engine does not only hand back the smallest number of really current web pages, it also shows – by far – the largest amount of outliers. As seen, the two most outdated pages in the complete research are also delivered by the Yahoo index.

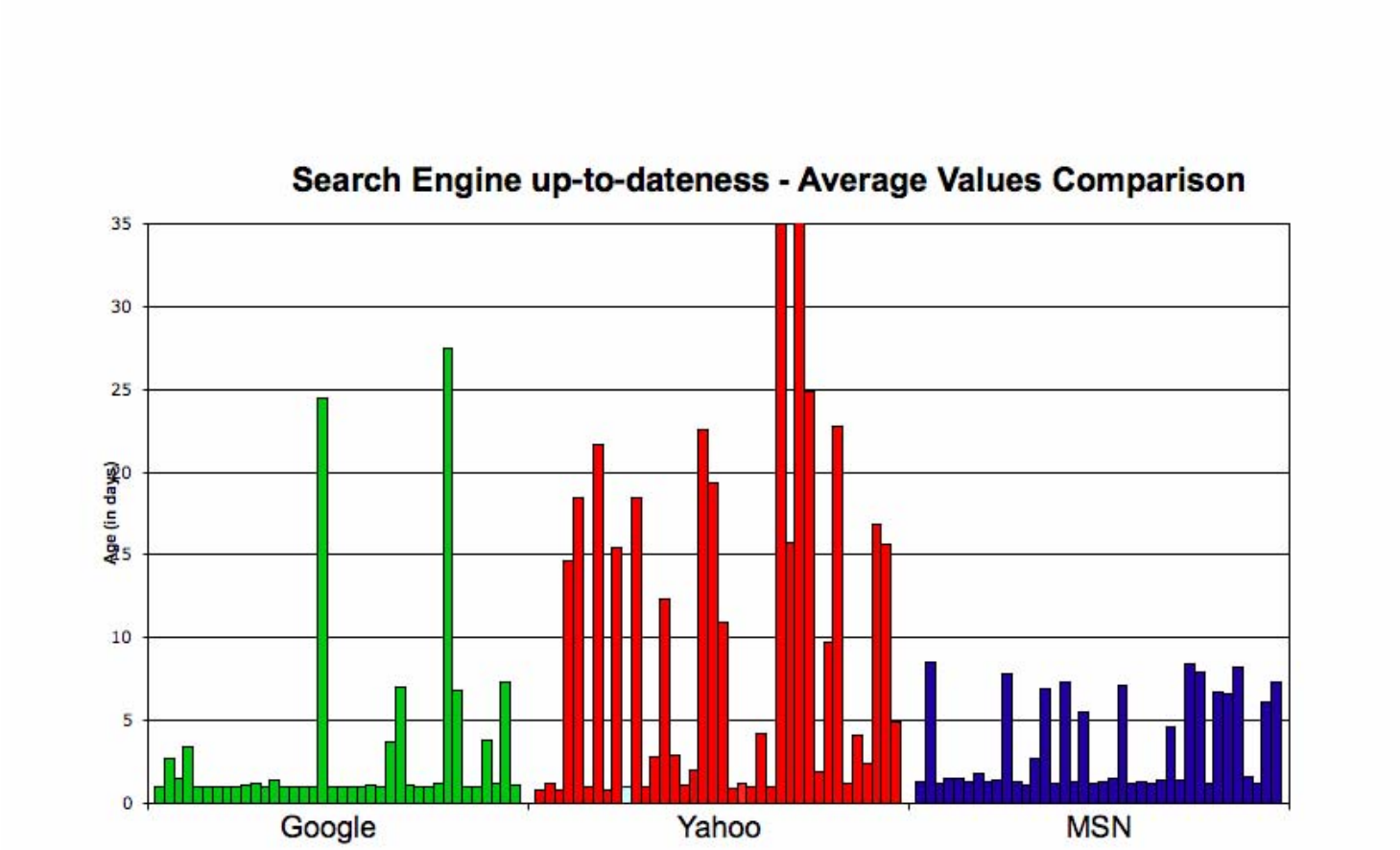

Fig. 2. Search engine up-to-dateness – Average values comparison

The following three graphs (fig. 3-5) show the frequency distribution of all records on a day-scale. Again, 0 and 1 day old pages were taken together to get a more reliable picture. If we take a closer look at this we can see that Google exceeds the other competitors by far with the largest amount of current web pages. It reaches a total number of 1291 records whereas MSN and Yahoo stay below 750. Nevertheless the oldest Google page in the sample data was 54 days old.

Generally speaking the pages in the MSN index are usually older. But - although keeping the limitations of our data set in mind - we can say that MSN seems to be the only search engine here that updates its index completely within a time-span of less than 20 days.

The Yahoo results are much more dispersed. Yahoo has the smallest amount of pages which are 0 or 1 day old. Additionally this search engine handed back the oldest page of our study, which was 62 days old. This graph clearly indicates that Yahoo has obvious problems in updating its index continuously.

Looking at the variance of the results for the individual engines, we find that MSN has the lowest variance with a standard deviation of 3, while Google reaches 6 and Yahoo even 10.

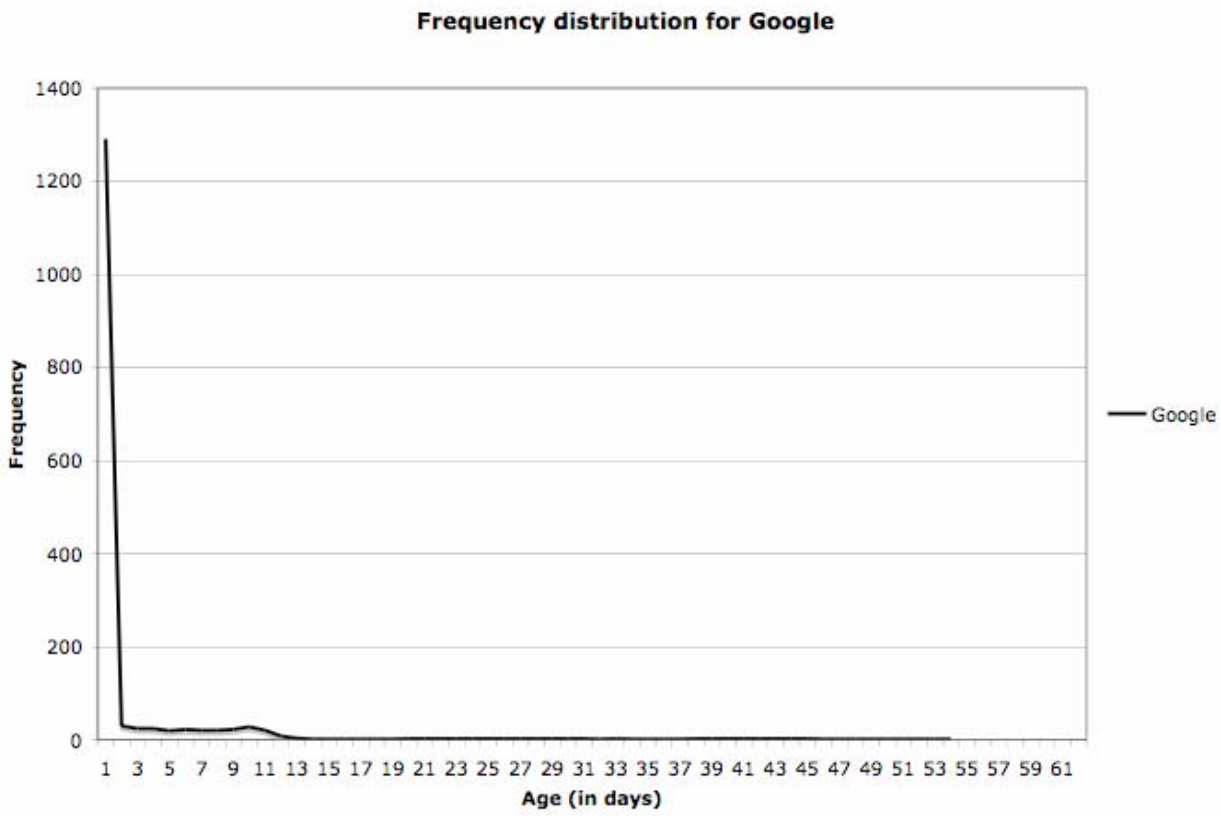

Fig. 3. Frequency distribution for Google

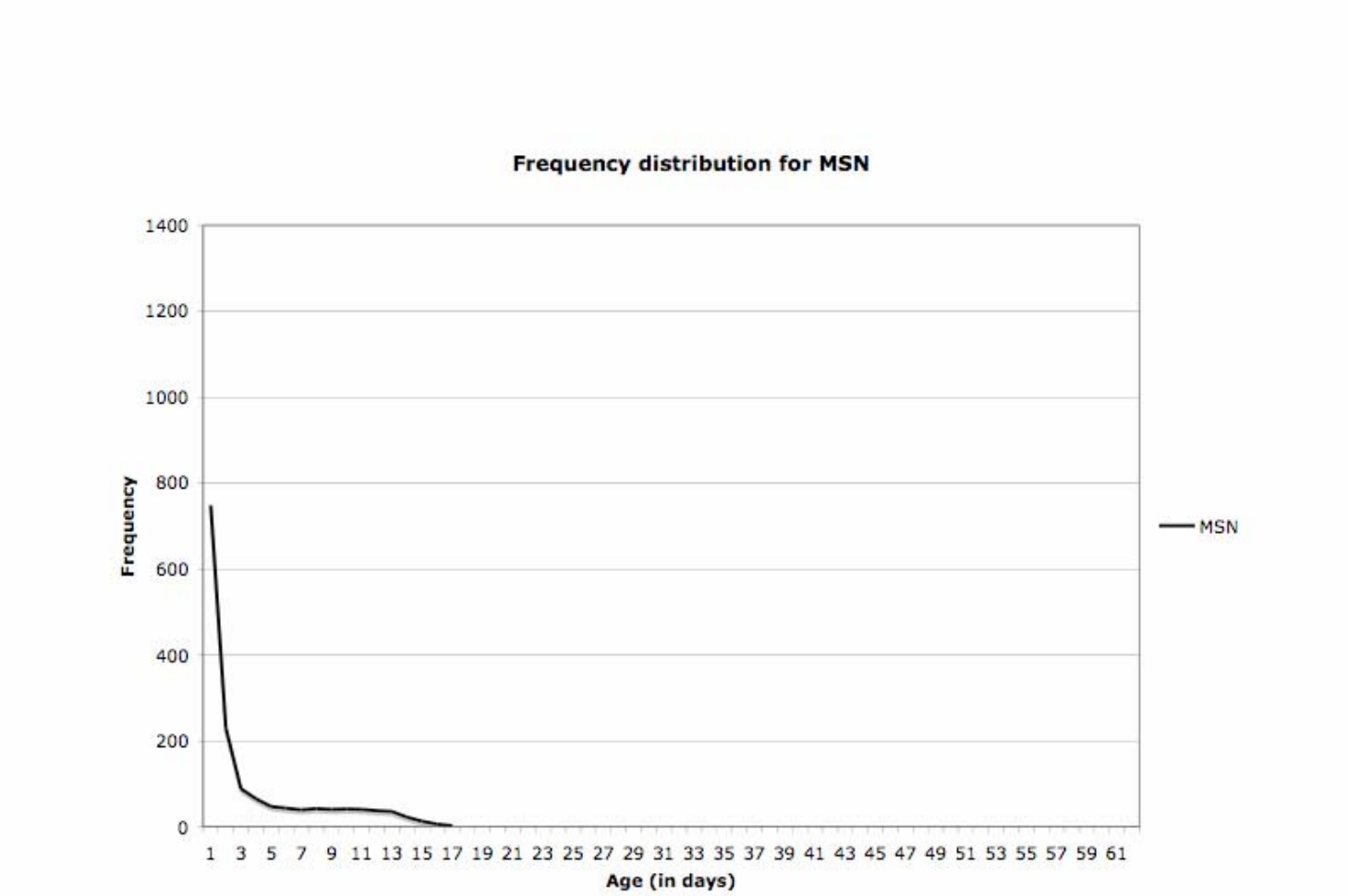

Fig. 4. Frequency distribution for MSN

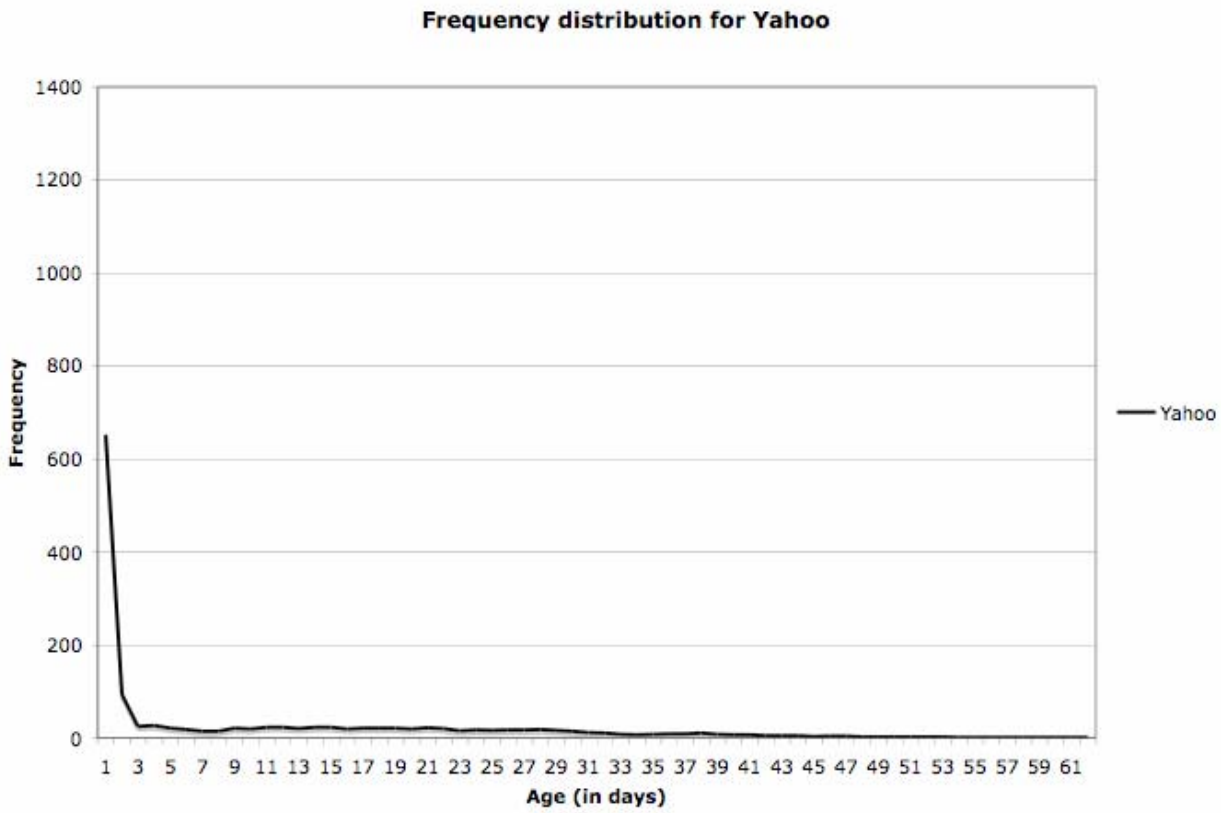

Fig. 5. Frequency distribution for Yahoo

5.2. Results for the different groups of websites

As described in chapter 5.1, we recognized that Google returned the freshest overall results. But what about the four groups which were formed in terms of content? Can we see any differences regarding the content and popularity of the website?

Therefore we will now compare the arithmetic means for all members of the four groups separately.

5.2.1 National news portals

National news portals are a group of very popular websites. Most of them are updated within very short time-spans, so it is extremely important that the items of the index referring to these sites are updated by the search engines very fast.

In this group, Google offers the best results (see figure 6), if we calculate an overall average for all nine members of the group. No page is older than 4 days. Nevertheless, MSN returns even better results for most members in the group, but falls down because of its bad results concerning http://www.swr.de, one of the group members. Finally, both reach a general average update interval of two days in this section. Yahoo also offers generally good results compared to the other sections, but has two extreme outliers (www.sportschau.de, www.wdr.de), which spoil the overall picture. Finally, it reaches a total indexing interval of 5 days for the national news portals.

Above all the national news portals are being updated by Google, MSN and Yahoo within an average frequency of three days.

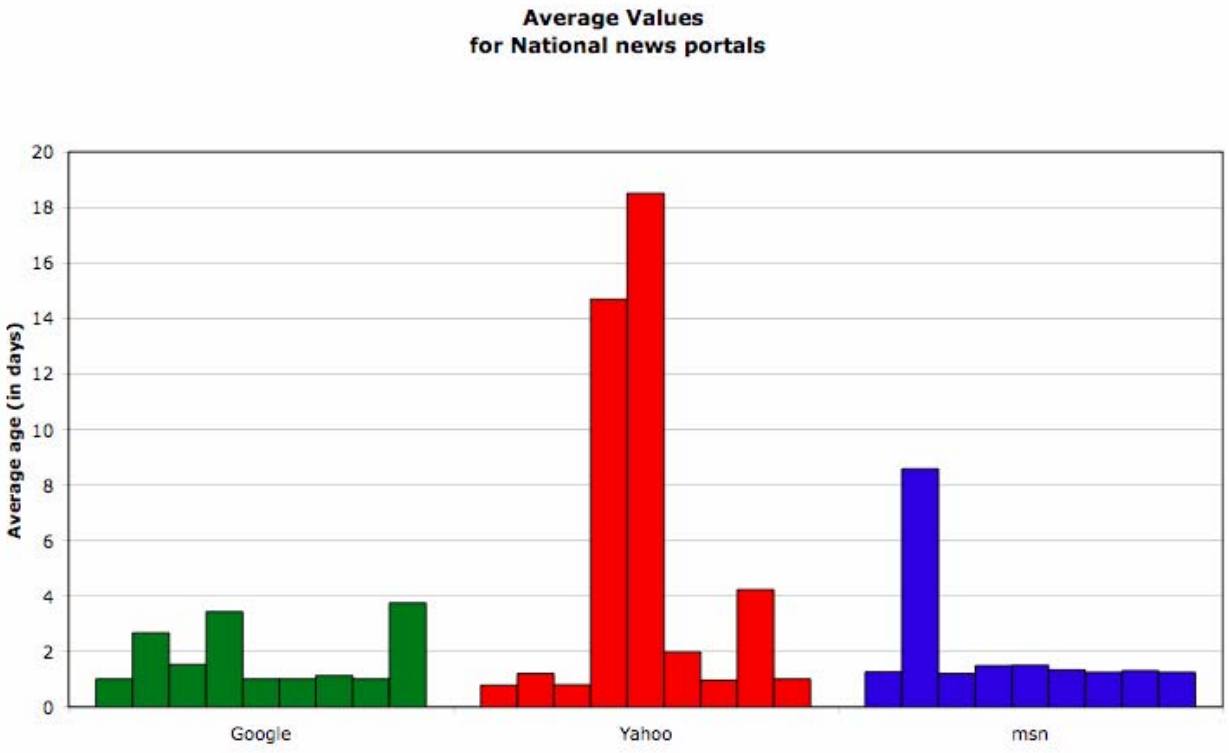

Fig. 6. Average values for national news portals

5.2.2 News portals with regional character

News portals with regional character are updated daily or even more often. The content on these sites addresses to local users, so these sites usually do not get the same traffic as the nationwide news portals in group 1. Does the update interval of the sites reflect this difference? The results are shown in figure 7. First of all, Google offers the best results again. Unlike to the first group, all pages have a continuous average age of one day. Google reaches better results here compared the nationwide ones. The hypothesis that up-to-dateness depends on popularity seems to be falsified.

Yahoo hands back extremely bad results with five of the nine pages older than 15 days and one outlier (www.ostfriesische-nachrichten.de) with an average age of 39 days.

MSN does not reach the results of the first group. Four of the nine web pages have an average more than 4 days of up-to-dateness.

This leads to a final average update frequency of six days for the group of regional news portals. Compared to the three days in the section of the nationwide sites, we finally can prove the hypothesis that popular websites are updated more frequently than comparable less popular ones.

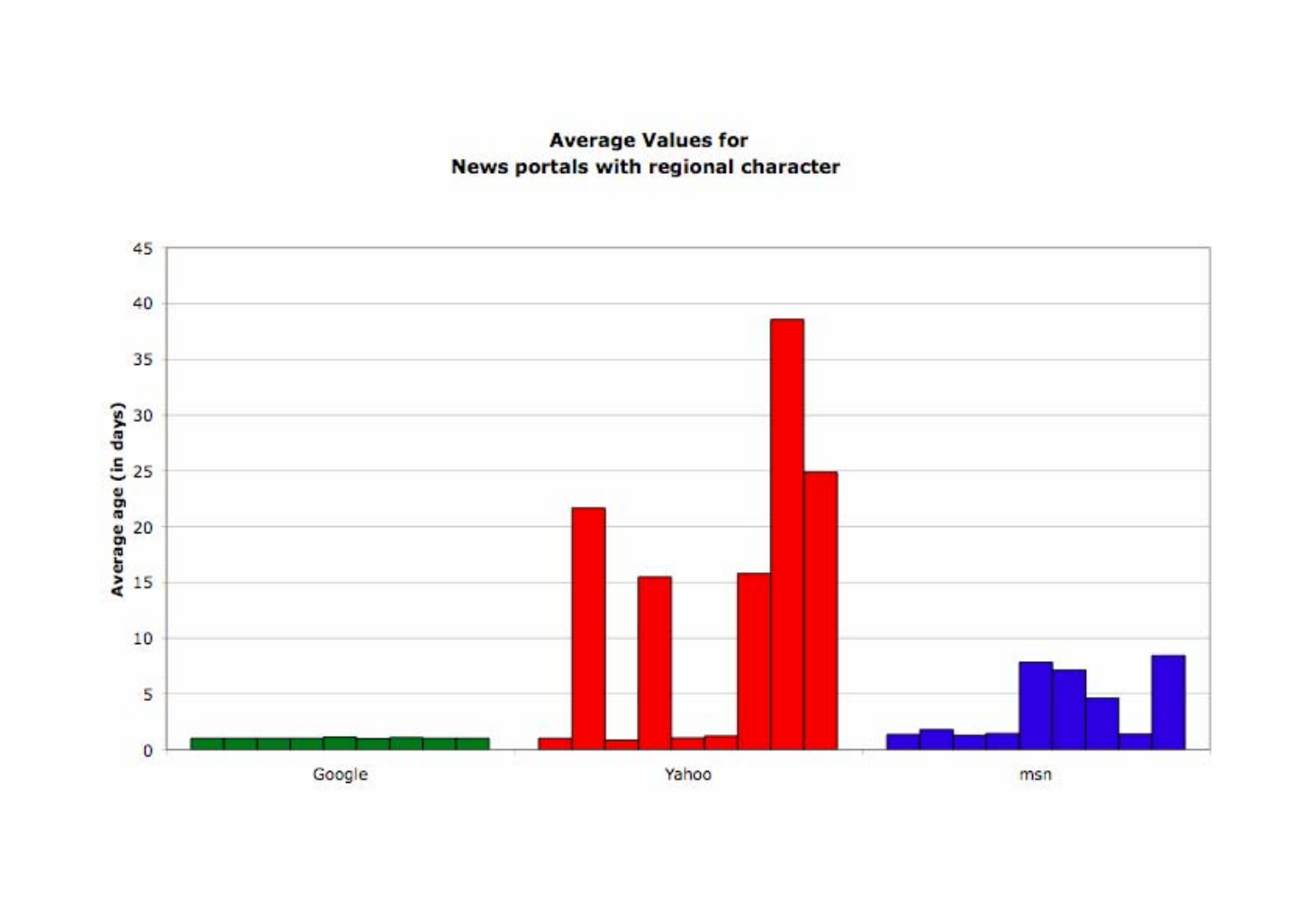

Fig. 7. Average values for news portals with regional character

5.2.3 Scientific oriented websites

Focusing on the three competitors we have another leader in this group (see figure 8). Google is loosing its top position mainly because of one member, the site www.diabetes.uni-duesseldorf.de. This outlier falsifies the results of Google. MSN instead gets continuously good results with an average age of four days for the whole group. So MSN is the leader in this group. Yahoo handed back the worst results again. Five of the ten sites have an average age of more than twelve days. One outlier has an mean average age of 35 days.

Generally, these separate results lead to a complete update average of six days for the section of scientific oriented websites. This means that these sites are updated with the same frequency than regional news portals.

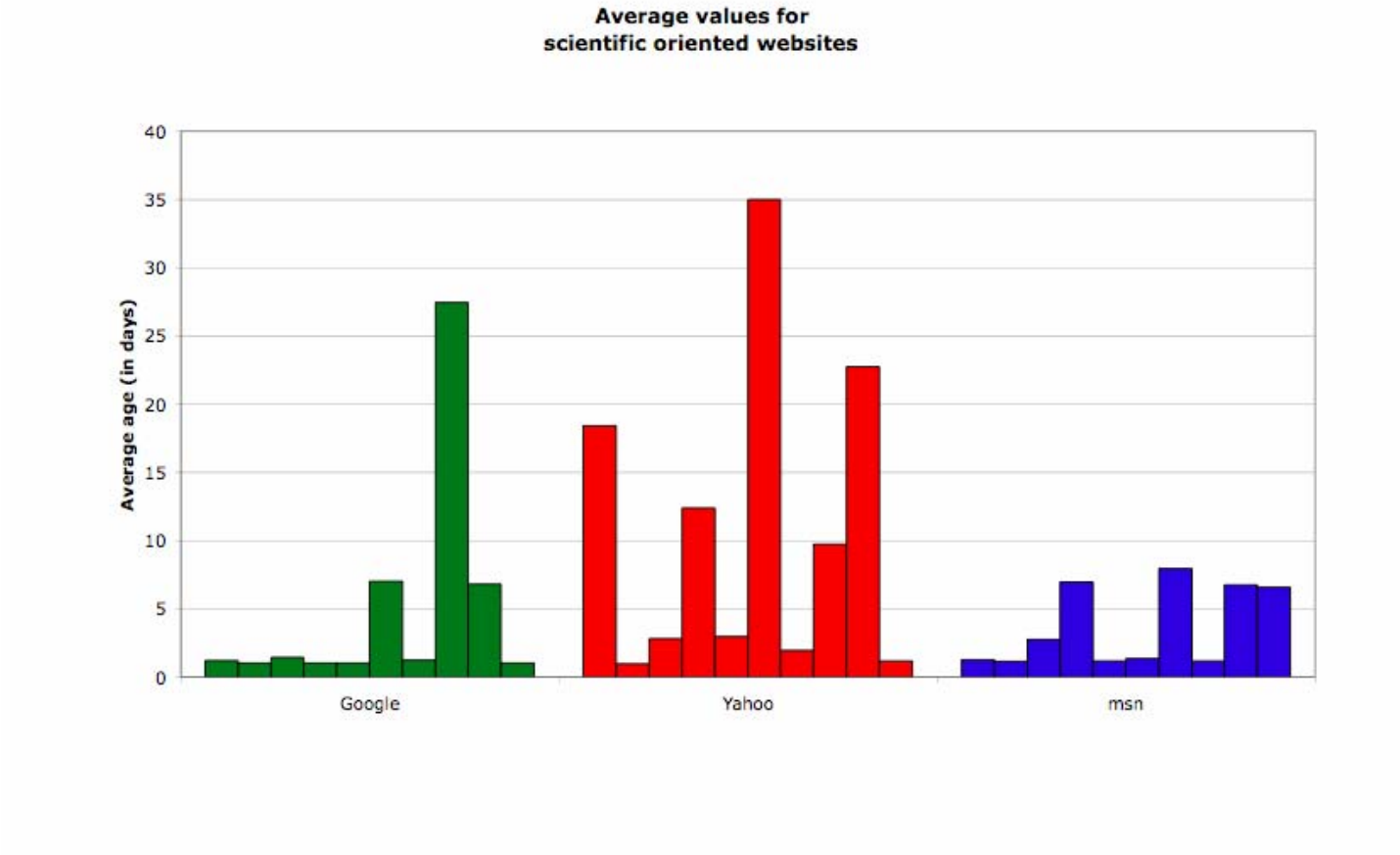

Fig. 8. Average values for scientific oriented websites

5.2.4 Special interest websites

The final group includes special interest websites with a clear focus on entertainment and hobby sites. The results for this group (figure 9) are generally quite similar to the results of the previous group. Google offers seven very good results with an average age of one day. But there exists an outlier with an mean average age of 25 days (www.musikmarkt.de). That is the reason why MSN is the winner in this group. No result of MSN exceeds the average age of eight days. Yahoo once more offers the worst results. Half of the ten pages are older than 11 days on average.

Also the general average update frequency reaches the same level when compared to the last group: Like the scientific oriented websites, the special interest websites are updated every six days in terms of average. This shows that there is no difference in the updating process concerning the content type of the website.

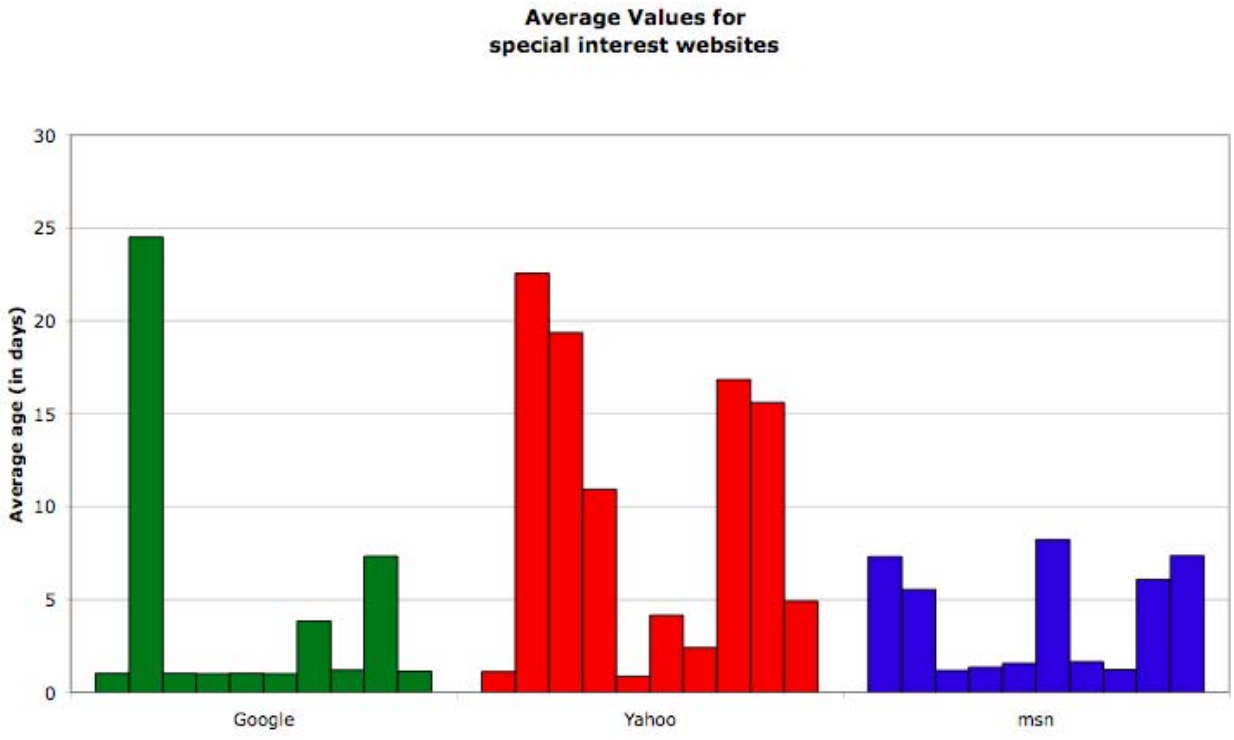

Fig. 9. Average values for special interest websites

Finally we compare the already presented arithmetic means of the four groups for the search engines separately (figure 10). It shows that Google got its best results in the group *News portals with regional character* followed by the group *National News portals*. The two other groups clearly fall behind.

In contrast to that MSN got its best results in the first group of *National news portals* followed by the other groups with an average result of four days. Again, we can see the continuity of the MSN index with comparable results in all four groups.

Yahoo shows shortcomings in all groups except for the *National News portals*. In this section, the search engine reaches a quite adequate result with an average of 5 days, whereas the other sections clearly fall behind. Yahoo will have to work on its updating process especially among these less popular websites if the company wants to reach the two competitors in the future.

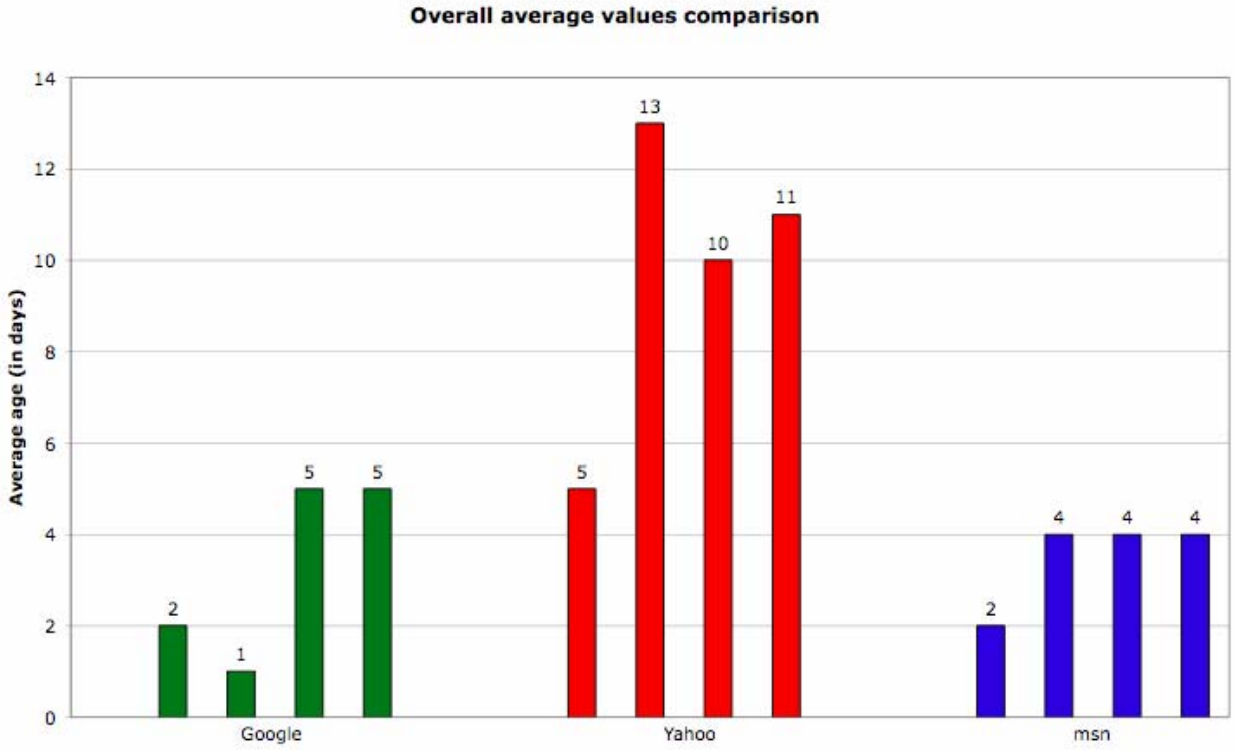

Fig. 10. Overall average values comparison

5.3 Indexing patterns

A specific focus of this research was to get more information on the patterns in which the search engines update their indices. The question was whether our results show specific intervals for every page, group of websites – or are there any patterns at all?

From our point of view importance should be attached to the observation of the indexing patterns. The knowledge about the reliability of the indexing patterns adds new aspects for an overall evaluation of the search engines. Only an engine guaranteeing constant updates of its index can maintain full credibility towards the users. Seen from this perspective, our results show a multifaceted image with different conclusions for every competitor.

5.3.1 Google

Google does not only hold the first position in our overall ranking. When focusing on the indexing patterns this engine is superior since the first look at the results is impressive: No other search engine updates so many sites as constantly fast as Google.

Selecting the fastest updated sites of our research we can see that 28 out of the 38 pages are updated by Google on a constant daily basis. This underlines the constant quality and up-to-dateness of the index.

But on the other hand we also find some outliers that contradict the prior findings: Two websites in our test were nearly completely forgotten by the Google robots. Their cache version became older and older and grew up to the final peak of 54 days (fig. 11). Six further pages showed noticeable breaks in their updating frequency (fig. 12). Between two days on both of which the current version was shown an outdated version was displayed. This phenomenon of sudden pattern breaks, as later seen even more obvious at Yahoo, implies technical problems in the automatic update process of the index. It reduces the general reliability of the Google results.

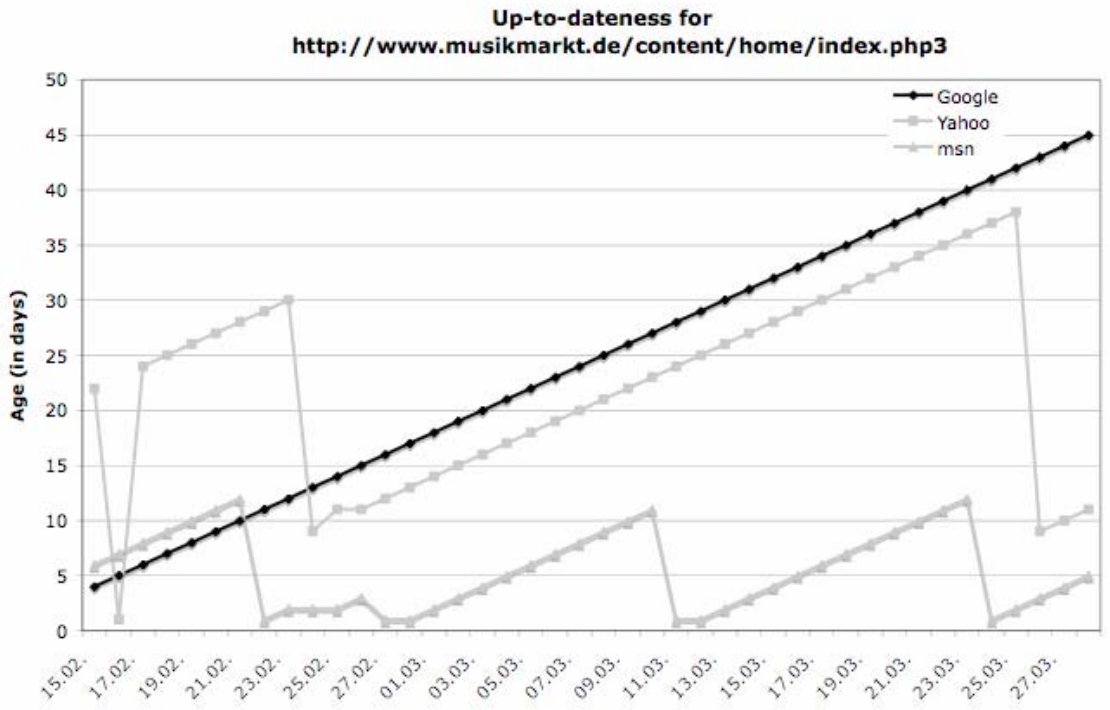

Fig. 11: Google forgets to update its index: Up-to-dateness for http://www.musikmarkt.de/content/home/index.php3

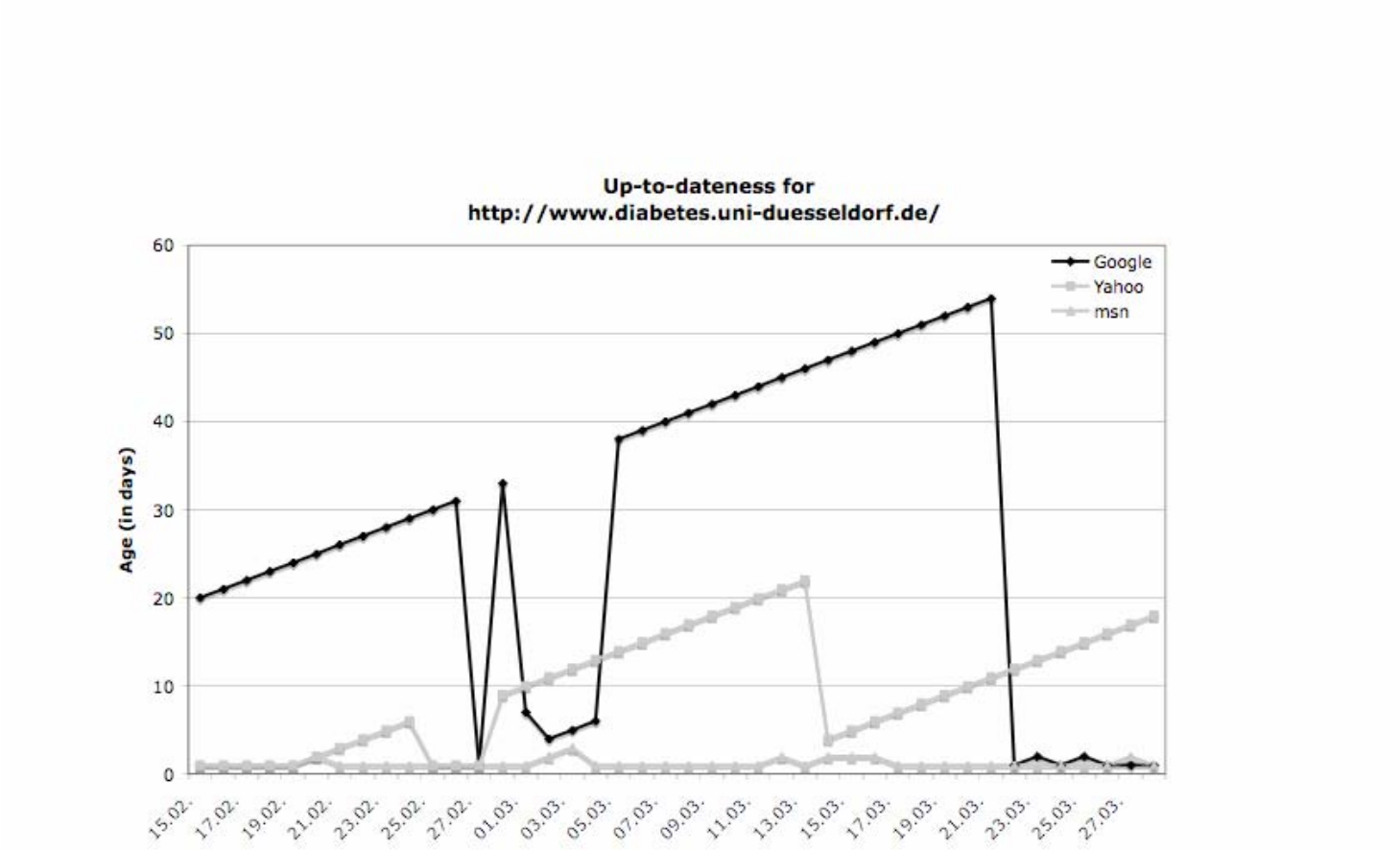

Fig. 12. Google - sudden pattern breaks: Up-to-dateness for http://www.diabetes.uni-duesseldorf.de

5.3.2 MSN

MSN in contrast is presenting a more constant picture, although it cannot reach the top positions of the Google frequency. Whereas Google can offer 20 constantly daily updated pages, MSN has not one page with an adequately frequent pattern. We already outlined that the number of absolutely current web pages reached here is definitely smaller than that of Google. MSN reached just 748 (48.01 percent), whereas Google had a percentage of more than 82 percent.

Taking a closer look at our data, we now can see the reason for that fact: Even the fastest updated MSN pages have some leaks in their daily update frequency. There is not one page at all which was constantly updated day

by day during the whole time-span of our research. But here the leaks, in contrast to Google, do not widen that extent. The phenomenon of inexplicable pattern breaks is hardly visible here. Instead, the results show a clear update pattern for all pages (fig. 13 shows a typical example): 19 pages are in the group for daily updates, another 14 pages have a constant update frequency of (about) 15 days. The rest has an update interval somewhere in between, but not a single page in the MSN index was ever completely out-dated. The graph again shows us that even the oldest results in our research were just 17 days old. Overall it can be said that MSN has not the fastest, but definitely the technically most reliable index.

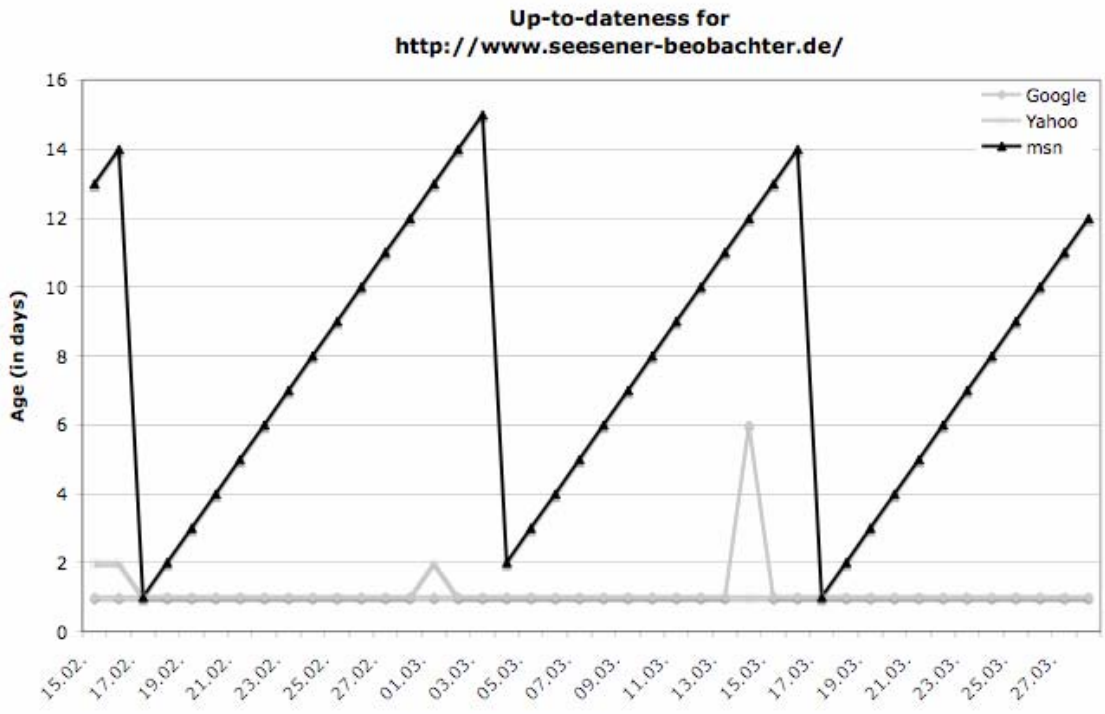

Fig. 13. Typical MSN update pattern (up-to-dateness for http://ww.seesener-beobachter.de)

5.3.3 Yahoo

Like in the overall results, the Yahoo index is behind its two competitors in terms of update frequency. Moreover, it must be said that Yahoo has the only index with *general* problems in its updating process.

The phenomenon of the sudden pattern breaks, as already mentioned, is a much more visible and common problem at Yahoo: 16 out of all 38 pages in our research show inexplicable breaks in its update frequency. They sometimes exceed more than 20 days.

An example for this problem is www.sportschau.de (fig. 14). It suddenly displays a 24 days old cache version on 21st February after six days with straight one-day-old versions before. On the next day, the out-dated version is again followed by a current version from the day before. But later it returns to the out-dated version, which now reaches an age of 26 days. From now on, the display in the Yahoo result list is constantly swapping between these two (and sometimes even more) versions from the cache.

This is a typical phenomenon for many Yahoo pages in this research. It seems to be part of a general problem in their updating policy. Among several versions in the cache, Yahoo cannot find the latest one.

These difficulties contradict the ambitions of Yahoo, which is trying to set up a fast indexing pattern like Google. From time to time this really works, but just for eight pages which were updated daily during the whole time-span of our research. Most of the other pages were affected by the technical problems to such an extent that one cannot rely on the up-to-dateness of Yahoo. The company has to work on that problem to be able to compete with the other engines in terms of index freshness.

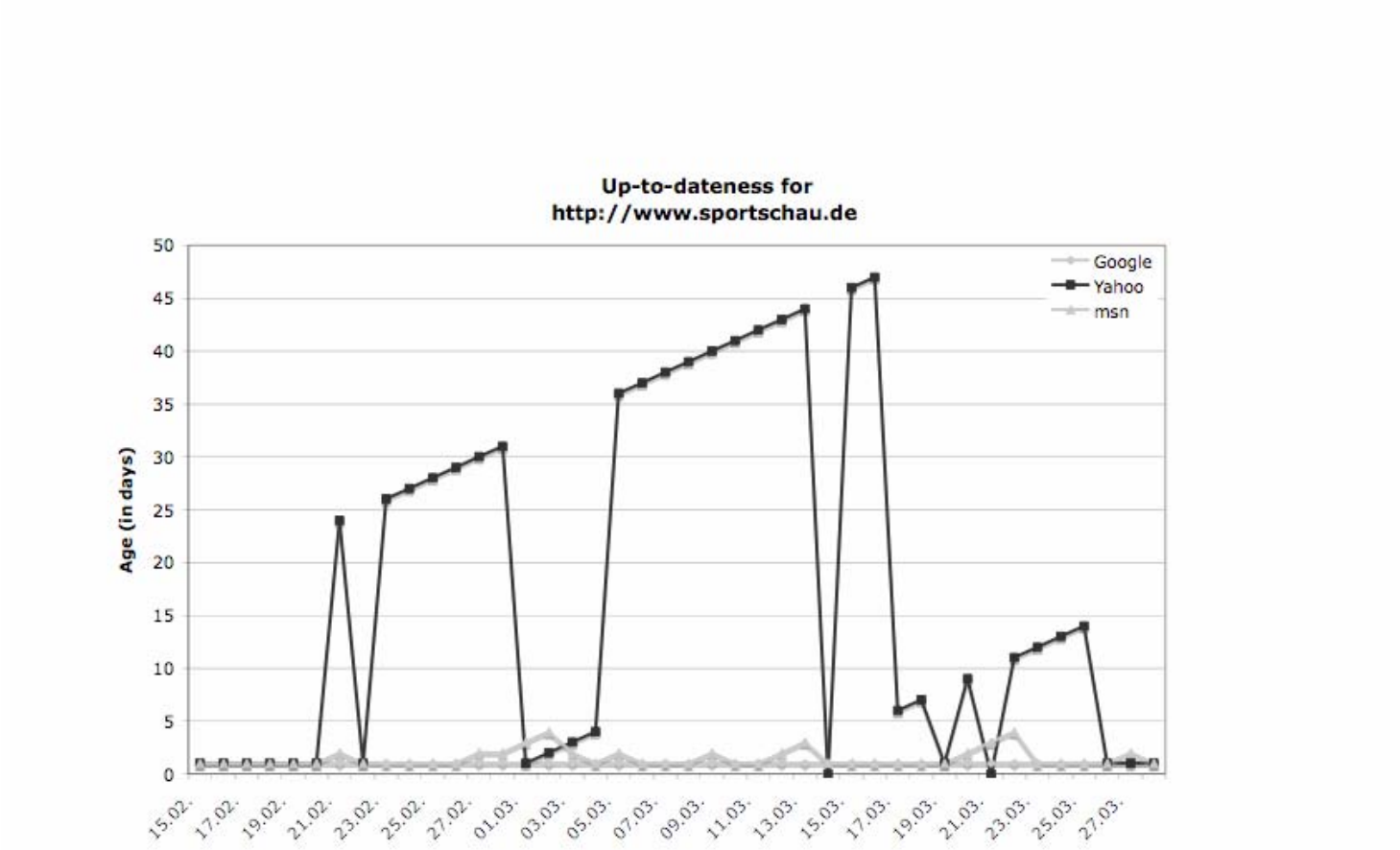

Fig. 14. Yahoo up-to-dateness for http://www.sportschau.de

6 Discussion and conclusions

Considering the results of this research we can say that only MSN seems to be able to update its whole index within a time-span of less than 20 days. Since our research only focussed on web pages that are updated on a daily basis, we cannot be too sure about this. On the basis of our findings we can conjecture 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 our findings Google is the fastest search engine in terms of index quality, because many of the sites were updated daily. In some cases there are outliers that were not updated within the whole time of our research or show some noticeable breaks in their updating frequency. In contrast to that MSN updates the index in a very clear frequency. Many of the sites were updated very constantly. Taking a closer look at the results of Yahoo, it can be said that this engine has the worst update policy. The graphs indicate a lot of peaks and breaks. All in all the company seems to update in a chaotic way.

It would be interesting to know on which criteria the search engines base their update policy. What is the reason why one web page is updated within a day and another within a week? We think, the more popular a website seems to be, the more often this site should be updated. But how does a search engine recognize how popular a website is? Does it count the number of visitors or the amount of links that refer to the site? Both methods seem possible since all search engines count backlinks and rate them due to their importance and also determine the click popularity by collecting user data via toolbars. Further research should combine data about a website such as size, backlinks in the search engines' databases and usage data with data about the update frequency of the search engines. Therefore, data from known websites should be used. The problem is that one needs internal data (e.g. usage data) from the website provider to do this. Because of the non-availability of such data for the given set of websites we omitted such analysis from our study.

Our research can only reveal findings for a set of representative web pages with a daily update. We therefore cannot say anything about the general update frequency of the search engines. But as our results show, even pages that are updated daily can reach a very low update frequency in Google as well as in Yahoo. From the results of our study, we can assume that MSN updates its whole index more frequently than its competitors. But none of the examined engines comes close to the ideal of an update frequency of just one day. It would be interesting to see whether the search engines will be able to improve the freshness of their databases or at least keep it even with growing indices. Therefore, our research has to be repeated and the results have to be compared to our current findings.

In terms of the groupings we made we have to admit that the groups are too small to give exact results. We assume that pages which are updated on a more frequent basis are indexed more frequently by the search engines. Looking at our comparison of science oriented pages and special interest pages, we can find no significant differences that would let us state that the search engine updates depend on content factors.

We finally cannot give a clear recommendation to which search engine should be used for a search on current content. With MSN, one can be sure that the indexed pages are at least 20 days old, but looking for pages updated only recently or newly published sites, one cannot be sure if these are already indexed by MSN. When using Google one can assume that most of the pages are updated within a very short time-span but cannot be sure whether there are some outliers which have not been updated for a long time. So the only clear recommendation we can give is not to use Yahoo when it comes to searching for current content. This search engine updates its

index in a rather chaotic way, so one can neither be sure how old the index actually is nor if a large proportion of the pages in the index was updated recently.

6. Acknowledgements

We would like to thank Anne Beckers and Anna Kalita for participating in the pretest of this study.

7. References

[1] A. Acharya, A., 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 (Patent Application US 2005/0071741 A1, 2005)

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

[3] N. Ford, D. Miller and N. Moss, Web search strategies and retrieval effectiveness: an empirical study, *Journal of Documentation* 58(1) (2002) 30-48

[4] R. Fries, W. Schweibenz, J. Strobel and P. Wiland, Was indexieren Suchmaschinen? Eine Untersuchung zu Indexierungsmechanismen von Suchmaschinen im World Wide Web, *BIT Online* 4(1) (2001) 49-56.

[5] J. Griesbaum, *Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de* (2004). Available at: http://informationr.net/ir/9-4/paper189.html (accessed 8 May 2005).

[6] J. Griesbaum, M. Rittberger and B. Bekavac, Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de. In: R. Hammwöhner, C. Wolff, C. Womser-Hacker (eds.), *Information und Mobilität. Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft* (UVK, Konstanz, 2002).

[7] S. Lawrence and C.L. Giles, Searching the World Wide Web, *Science* 280 (1998) 98-100.

[8] S. Lawrence and C.L. Giles: Accessibility of information on the web. *Nature* 400(8) (1999) 107-109.

[9] H. Leighton and J. Srivastava, First 20 Precision among World Wide Web Search Services (Search Engines), *Journal of the American Society for Information Science* 50(10) (1999) 870-881.

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

[11] L. Lo Grasso and H. Wahlig, Google und seine Suchparameter: Eine Top 20-Precision Analyse anhand repräsentativ ausgewählter Anfragen. *Information Wissenschaft und Praxis* 56(2) (2005) 77-86.

[12] M. Machill and C. Welp (eds.), *Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen* (Verlag Bertelsmann Stiftung, Gütersloh, 2003).

[13] M. Machill, D. Lewandowski and S. Karzauninkat, Journalistische Aktualität im Internet. Ein Experiment mit den "News-Suchfunktionen" von Suchmaschinen. In: M. Machill and N. Schneider (eds.), *Suchmaschinen: Eine Herausforderung für die Medienpolitik,* (Vistas, Berlin, 2005).

[14] A. Mowshowitz and A. Kawaguchi, Assessing bias in search engines, *Information Processing & Management* 38(1) (2001) 141-156.

[15] G. Notess, *Search Engine Statistics: Freshness Showdown [Data from 17 May 2003]* (2003). Available at: http://www.searchengineshowdown.com/stats/freshness.shtml (accessed 17 April 2005).

[16] G. Notess, *Search Engine Statistics: Freshness Showdown [Data from 20 October 2002]* (2002). Available at: http://www.searchengineshowdown.com/stats/0210freshness.shtml (accessed 17 April 2005).

[17] G. Notess, *Search Engine Statistics: Freshness Showdown [Data from4 April 2002]* (2002). Available at: http://www.searchengineshowdown.com/stats/0204freshness.shtml (accessed 17 April 2005).

[18] G. Notess, *Search Engine Statistics: Freshness Showdown [Data from 7 March 2002]* (2002). Available at: http://www.searchengineshowdown.com/stats/0203freshness.shtml (accessed 17 April 2005).

[19] G. Notess, *Search Engine Statistics: Freshness Showdown [Data from 13 August 2001]* (2001). Available at: http://www.searchengineshowdown.com/stats/0108freshness.shtml (accessed 17 April 2005).

[20] A. Ntoulas, J. Cho and C. Olston, What's New on the Web? The Evolution of the Web from a Search Engine Perspective (2004). In: *Proceedings of the Thirteenth WWW Conference, New York, USA.* http://oak.cs.ucla.edu/~ntoulas/pubs/ntoulas_new.pdf (accessed 8 May 2005).

[21] A. Singhal, and M. Kaszkiel, A Case Study in Web Search using TREC Algorithms. In: *Tenth World Wide Web Conference 2001: Proceedings of the 10th World Wide Web Conference* (ACM Press, New York, 2001).

[22] D. Sullivan: Nielsen Net Ratings Search Engine Ratings, Searchenginewatch.com. http://searchenginewatch.com/reports/article.php/2156451 (accessed 22 April 2005).

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

[24] C. Wolff, Effektivität von Recherchen im WWW: Vergleichende Evaluierung von Such- und Metasuchmaschinen. In: G. Knorz and R. Kuhlen (eds.), *Informationskompetenz - Basiskompetenz in der Informationsgesellschaft, Proceedings des 7. Internationalen Symposiums für Informationswissenschaft* (UVK, Konstanz, 2000).

Appendix A: List of sites used for this study

1. National news portals

http://www.web.de: One of the biggest German portals with a huge website catalogue, free mail, news updates and other free services.

http://www.swr.de/nachrichten: News service of the *Süddeutscher Rundfunk* (Southern German Broadcasting Company), public broadcasting service for approx. 14 million Germans.

http://www.faz.net/s/homepage: Front page of the *Frankfurter Allgemeine Zeitung* (FAZ), one of the leading national daily German newspapers.

http://www.wdr.de/radio/nachrichten: Contains a list with the current radio news from the *Westdeutscher Rundfunk* (Western German Broadcasting Company), public broadcasting service for approx. 18 million Germans.

http://www.sportschau.de: Homepage of the most important sports show on German television and one of the leading sources for sports information on the German web.

http://www.dradio.de/aod/html: Information page of the *Deutschlandradio*, a public and nationwide German radio station.

http://www.reuters.de: Homepage of the German branch of the international news agency *Reuters*.

http://www.mdr.de/nachrichten: News service of the *Mitteldeutscher Rundfunk* (Central German Broadcasting Company), public broadcasting service for approx. 9 million Germans.

http://www.diepresse.com: Home of the leading Austrian daily newspaper *Die Presse*.

2. News portals with regional character

http://www.chiemgau-online.de: News portal for the Chiemgau region in South-east Bavaria.

http://www.nordclick.de: Common news portal from daily papers from Schleswig-Holstein, a region in Northern Germany.

http://www.rp-online.de/public/home/nachrichten: Front page of the *Rheinische Post*, a leading daily newspaper in Duesseldorf, the capital of the German region North Rhine-Westphalia.

http://www.fraenkischer-tag.de/nachrichten: Internet base of the *Fraenkischer Tag*, a daily newspaper offered in some parts of Franconia (part of the German region of Bavaria).

http://www.seesener-beobachter: Local news from Seesen, a medium-sized town in the German region of Lower Saxony.

http://www.merkur-online.de: Web service of the *Münchner Merkur*, a leading daily newspapers in the Bavarian capital Munich.

http://www.stuttgarter-zeitung.de/stz/page/detail.php/13: Weather forecast of the *Stuttgarter Zeitung*, a newspaper for the capital of the German region of Baden-Württemberg.

http://www.ostfriesische-nachrichten.de/neu/index.asp: Homepage of a local daily newspaper *Ostfriesische Nachrichten* in East Frisia (part of Lower Saxony in North-West Germany).

http://www.bbv-net.de/public/home/nachrichten: Homepage of the *Bocholt-Borkener Volksblatt*, a German local daily newspaper in Western Westphalia.

3. Scientific oriented websites

http://www.idw-online.de/pages/de: Front page of the *Informationsdienst Wissenschaft* (Academic Information Service), a big nationwide provider for current academic news in Germany.

http://www.uni-duesseldorf.de: Homepage of the *Heinrich-Heine-University* in Duesseldorf.

http://de.wikipedia.org/wiki/Hauptseite: Front page of the German branch of the *Wikipedia* project, an open-source encyclopaedia on the web.

http://www.uro.de: German news portal for the latest scientific developments in the urology.

http://www.pro-physik.de: Online-initiative of the German Physics Association with updates on current research in physics.

http://www.medianrw.de/kurznachrichten/index.php: News listing for experts of the media business. The page is published by the regional government of North Rhine-Westphalia.

http://www.aerztezeitung.de: Online-Version of the *Ärztezeitung*, a magazine especially set up for doctors.

http://www.diabetes.uni-duesseldorf.de: Web portal of the German Diabetes Centre, located at the University of Duesseldorf.

http://www.aerztlichepraxis.de/aktuell/nachrichten: Online news service with clinical information for doctors and patients.

http://www.stmwfk.bayern.de: Homepage of the Ministry for Science, Research and Arts of the provincial government in Bavaria.

Special interest websites

http://www.camping-channel.com: First hand information base for campers.

http://www.musikmarkt.de/content/home/index.php3: Current news from the *Musikmarkt*, the leading German magazine in the music business.

http://www.biograph-online.de/heute.php: Overview of the daily movie programme in Duesseldorf.

http://www.konsolen.net: News platform for people who are interested in video games.

http://www.finanznachrichten.de: Latest news from the stock markets.

http://www.daserste.de/wwiewissen: Homepage of the German TV show *W wie Wissen*, explaining scientific contexts to the lay mind.

http://www.wdr.de/tv/frautv: Homepage of a German TV show for women called *frauTV*.

http://www.la-palma-aktuell.de/cc/news.php: News service in German from the island of Majorca.

http://www.golfparadise.com: Information portal for golf sports.

http://www.liebesalarm.de: Online flirt community.