Date restricted queries in web search engines[1]

Dirk Lewandowski

Department of Information Science, Heinrich-Heine-University Duesseldorf,
Universitaetsstrasse 1, D-40225 Duesseldorf, Germany.
E-Mail : dirk.lewandowski@uni-duesseldorf.de

Search engines usually offer a date restricted search on their advanced search pages. But determining the actual update of a web page is not without problems. We conduct a study testing date restricted queries on the search engines Google, Teoma and Yahoo!. We find that these searches fail to work properly in the examined engines. We discuss implications of this for further research and search engine development.

Introduction

Some studies have shown that search engine users only seldom use the advanced search options, which all major engines offer. For a recent discussion of these findings, see for example Spink (2003) or Machill et al. (2003). Lewandowski (2004) gives an overview of the advanced features of the search engines Teoma, Google, AllTheWeb, Alta Vista, HotBot and Fireball.

Asking why advanced search features are seldom used one finds explanations like that the users do not want to invest time and brain power in either expressing their information need or formulating their queries (Machill et al. 2003, 169). The same study also found that only about half of search engine users know about Boolean operators and only 20 percent of search engine users use them frequently. Advanced search interfaces perform only slightly better. 59 percent of users know about them and only 14 percent use them regularly (Machill et al. 2003, 167-169).

This explains why Boolean operators are not used more often but it fails to explain why evident features like date restricted searching are not used at all. Date restricted queries seem easy to create and no or little search experience is required to understand them. Users might want to conduct a date restricted search when looking for recent documents or “hot topics”. In case of hot topics, there are often numerous documents that deal with older and not with the current events. For example, a user searching for Michael Jackson might be interested in the “hot topic” of the allegations against Jackson regarding child abuse. However, a search engine would list older documents first that might actually deal with the desired topic but not with the recent allegations.

[1] to appear in: Online Information Review 28(2004)6

There are many cases where date restrictions are appropriate and desired. If date restricted queries worked in web search engines it would be suitable to add a date range search box on the simple search pages.

However, there are some difficulties with date restricted queries on web search engines. These will be explained in the first part of this article. On the other hand some authors complain about the bad performance of these kinds of queries on the popular search engines (e.g. Price and Tyburski, 2003). There is no systematic analysis on this topic, though. For this reason, the present study focuses on a comparison between popular web search engines in regard to their ability to deliver relevant results from date restricted queries. In a conclusion we discuss some implications in order to improve the results.

Problems with date searching

In conventional information retrieval systems, searching with date qualifiers retrieves documents published or created on, before or after a specific date as well as documents created within a certain date range. We define the date of a document as the date when its contents were last changed. Other modifications such as changes in layout or the update of copyright information is not considered as a document update. Therefore, those changes do not affect the document date.

In HTML there is only a copyright meta tag to define the document date. This is rarely used and search engines have to rely on other information in determining the document date. There are mainly four possibilities to achieve this, which are

- selecting the date provided by the server on which the document is hosted
- using the date of first discovery by the search engine
- using the date quoted in the meta data of the document
- using date information given in the contents of the web page

Using the date provided by the server causes the problem that the server just gives the date of the last update of the *file*, which does not mean that the contents of the file have been changed. Our definition of the document date does not count such changes as date changes. In addition, web sites that generate their documents with content management systems often produce files on the fly and the current date is sent back. Each time the file is requested from the server a new date is generated. Another problem occurs when the date provided by the server is set incorrectly. In these cases it is impossible to get the actual date of a document from server generated data.

The date of first discovery by the search engine can help determine the actual revision date of the document. Problems with this approach appear for all documents that are older than the search engine itself and have not been changed since then. Only the date of the documents first being crawled by the search engines may be assigned to it. Further problems occur when the index size of the search engine is limited (which it usually is) and the index should be expanded.

Date information within the meta tags seems to be sufficient for determining the document date. Standard meta data sets support the date qualifier as do specialized sets such as Dublin Core. In addition, meta data sets define the date format.

However, a test run prior to this study showed that only a very small number of web sites use meta data in general and the date qualifier in particular. We found out that the date qualifier was used in only four documents out of about 500. Therefore, we ignored meta data in our further investigations.

The fourth possibility to determine the date of a document on the web is to extract the date directly from the contents of the document. The text often includes date specifications that can be used. To extract these date specifications one can look for standard date formats such as dd/mm/yyyy and use them for further analysis. There are various date formats (e.g. European vs. US) but this seems to be a minor problem. Date specifications usually occur in certain parts of the documents, e.g. at the beginning or end of the text or in the right upper corner. This allows automatic extraction systems to find the actual update information and to omit date information quoted in the text that is only relevant to the contents of the document.

In some cases date information given on the page such as "last updated on…" is generated on the fly so that each time the document is retrieved the date is set to the actual time of day. In order to cope with this problem one needs to compare the new and the old version of the retrieved document.

Today's search engines analyse the date information provided by the server while some of them use the date of first discovery. Meta data is not generally used for its limited availability. Search engines do not use the date information that is included in the documents. This is confirmed by our findings.

On the advanced search pages of the search engines the date restricted search is usually labelled with "return web pages updated within" or something alike. So the user is mislead; he or she will get texts from the chosen period and not just for example all those files whose layout was changed in the stated time span.

Objectives of the study

After reviewing the fundamental problems of date restricted searches we wanted to test whether search engines are able to correctly determine the dates of retrieved documents. For this purpose we randomly chose 50 queries and sent them to three different search engines. One time all queries were sent without date restriction, and another time they were sent with a restriction to documents updated during the last six months. This particular time span was chosen because all search engines reviewed in this study support it. Google and Yahoo! do not support date range queries constructed by the user.

All tests were conducted on April 4, 2004.

The first 20 hits were rated according to their up-to-dateness, which we define as the fraction of documents that have been updated during the last six months. We wanted to determine whether a date restricted search is worthwhile for the user. A date restricted search seems worthwhile for the user when the engine displays only those documents that have indeed been updated within the defined time span. Older documents should be excluded from the results pages. Our second research question was which search engine would be the best to conduct date restricted searches.

Our study can reveal results of search engine tests for a certain time only. To cope with the fast changing nature of the web and to see whether the search engines show improvements in providing date information, we plan to execute identical test runs at different periods of time.

In our study we checked the retrieved web pages for update information on the page. When we found an update information it was taken down to be used for our analysis. When we did not find any update information or when the information was not explicit we omitted the hit from further analysis.

Method

Choice of search engines

In our study we examined the search engines Google, Yahoo! and Teoma. These are the most common engines with the worldwide largest indexes (see Sullivan, 2003). The once important search engines Alta Vista and AlltheWeb do not have their own indexes anymore and are therefore omitted from this study. They now use the database of their new owner, Yahoo!. The Yahoo! database itself should be seen as the successor to the Inktomi database. This is why we omitted the "classic" Inktomi search engines like HotBot from this study. Neither did we include specialized search engines that have indexes for one subject or one language area only.

Choice of queries

We wanted our test queries to be random. We used the "Live-Suche" of German search engine Fireball that shows queries actually conducted by users. Thus we made sure that our queries had been chosen at random and that they reflected real information needs. Because Fireball mainly serves German queries, we conducted the queries in German. The queries mainly reflect the usual interests obtained in search engine query logs (Spink et al., 2002) covering, for example, commerce, travel, people, computers, and entertainment. The queries were generally kept short. 25 queries consisted of only one search term each, 11 queries of two terms, 9 queries of three terms, and only 5 queries consisted of four or more terms. Some queries contained umlauts, but our experience shows that this no longer is a problem for search engines.

The queries for our test set were retrieved on March, 15 2004. In order to prepare these for our test, we omitted all pictures index queries and those queries applied to the international index provided by a partner search engine (both are marked by Fireball). We also omitted queries that obviously showed a pornographic interest. Finally, all duplicate queries were excluded.

Using this method, we chose 50 queries for further investigation. We also kept some extra queries for those cases where queries would produce zero results.

Test setting

The 50 chosen queries were sent to the different search engines. We analysed the first 20 hits on the results pages first in the simple search, then in a date restricted search mode to retrieve only documents that were less than six months old.

We used the German language interfaces of the search engines. All standard settings were retained so that documents in every language were retrieved. An exception was Yahoo.de where we adjusted the settings to “worldwide” because this engine retrieves only German language results by default.

We did not check the results for relevancy. The only criterion we considered was that of available date information within the document. We considered all file types but as to our queries we found HTML and PDF files only.

“Dead links” on the results pages were ignored. We continued our investigation until we had examined 20 documents. Paid results (“sponsored listings” etc.) which are placed on top or on the side of the results were omitted.

Noticeable problems with several search engines

As already shown by the pre-test all date restricted searches were without effect when using the advanced search form on Google. That meant, it didn’t make any difference whether the date restriction was used or not. The results (and their order) were quite the same. This bug existed for at least six months but seems to be eliminated by now. In our study we had to use the *date range:* operator that uses Julian dates.

Appraisal of the update information

The tests were conducted by five graduate students who were introduced to the study’s goals and to some rules regarding date information characteristics. They were asked to search for date information in the retrieved documents. When update information was found it should be noted on the data entry form. We applied the following rules:

- When explicit update information was found within the text, it was written down. For example, this kind of update information could be found by selecting an indicator such as “last modified...” Some text types such as press releases or newspaper articles, which usually contain date information, could be found that way.

  But some pages contain update information that is generated automatically. This update information does not count as an actual update. We also omitted documents that contained the actual time at which our investigation was conducted. In some cases we could identify automatically generated update information based on the text itself. These pages were omitted, too. Documents that contained automatically generated update information were counted separately.
- Copyright notices within documents usually quote a year instead of the exact date. In many cases these copyright notices are generated automatically and therefore are the same for all pages of a web site. Usually the current year is used for all pages. Consequently, we did not consider copyright notices with a year specified by 2003 or 2004 but saw a specification of 2002 or earlier as a sign of out-of-dateness of the page and omitted it from further investigation.
- Sometimes we found date specifications lying in the future. These were ignored.
- The persons conducting the tests were asked to carefully distinguish between European and US date style.

We found that 28 to 33 percent of the retrieved web pages contain date specifications (see tables 1 and 2). Neither the differences between results from simple search and date restricted search nor the differences between individual search engines are statistically significant. With about 30 percent of all documents containing date specifications we had a sufficient amount of documents for further tests.

Table 1: Ratio of web pages with update information within the indexes of search engines (simple search)

Search engine	Number of hits for the 50 test queries*	Number of pages containing update information	Ratio of pages containing update information (in percent)
Teoma	933	313	33.55
Google	978	308	31.49
Yahoo!	979	296	30.23

* For each query the first 20 hits were examined. Therefore the maximum number of hits was 1,000 per search engine. Some queries produced less than 20 hits, which reduced the number of total hits per engine.

Table 2: Ratio of web pages with update information within the indexes of search engines (date restricted search: six months)

Search engine	Number of hits for the 50 test queries*	Number of pages containing update information	Ratio of pages containing update information (in percent)
Teoma	933	308	33.01
Google	971	279	28.73
Yahoo!	972	284	29.22

* For each query the first 20 hits were examined. Therefore the maximum number of hits was 1,000 per search engine. Some queries produced less than 20 hits, which reduced the number of total hits per engine.

Results

Up-to-dateness of the retrieved documents

We measured the amount of documents from the top 20 list that actually were updated within the last six months. We define the fraction of these documents among the amount of all documents as up-to-dateness rate. We calculated the corresponding contingents of documents retrieved by the simple search as well as by the date restricted search.

Teoma does not find a larger amount of current documents by a date restricted search than by the simple search. Teoma finds the lowest amount of up-to-date documents as well. In the simple search Yahoo! gets an up-to-dateness rate of 40.5 percent, where Google gets 48.7 percent. Thus, in Google nearly every second document is a "current document" according to our definition without even using date restricted searching.

When we restrict our search to documents from the last six months, Yahoo! gets an up-to-dateness rate of 54.2 percent and Google a rate of as high as 59.5 percent. Having said so, Google, which proved to be the best search engine in this test, fails in about 40 percent of all documents.

Looking at the enhancements in up-to-dateness, Yahoo! gets the best result. While Google gets an overall better result with 59.5 percent of correctly assigned hits, Yahoo! shows an increase of 33.7 percent. It seems that Google generally favours newer documents and frequently updated documents respectively.

Looking at individual documents rather than at the sum of all documents, one finds that the distribution of the up-to-dateness rate differs depending on the search query (see figs. 1-3). None of the search engines is able to only produce high or medium up-to-dateness rates for all queries. However, Google as well as Teoma produce an up-to-dateness rate of 100 percent significantly more often than Yahoo! does. On the other hand, both engines more often get a value lower than ten percent.

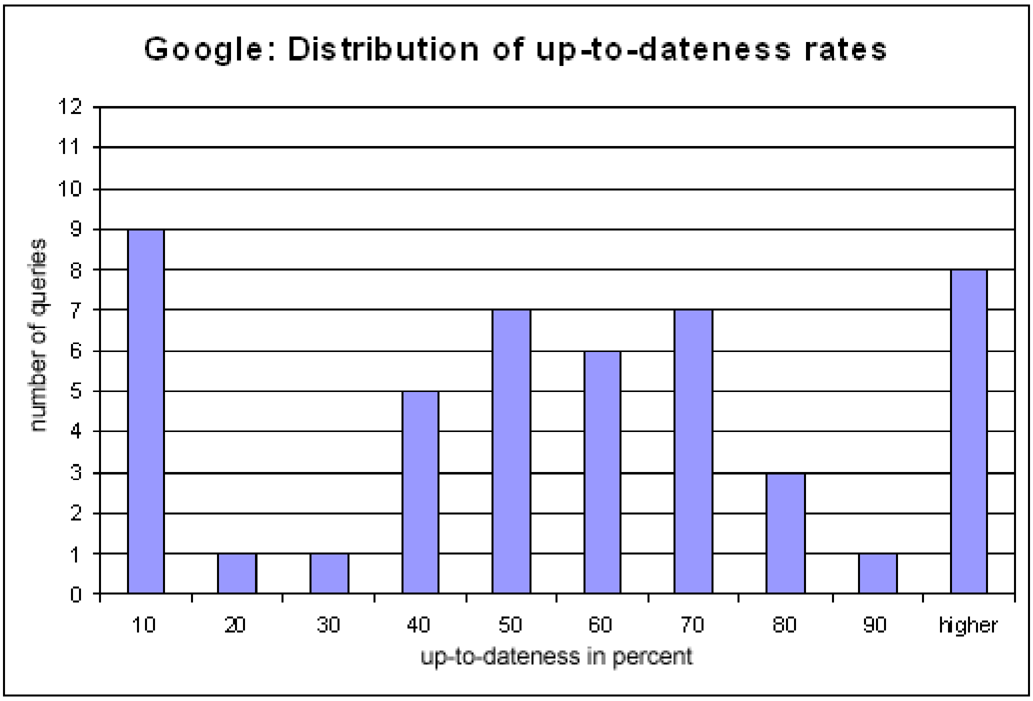

Figure 1: Distribution of up-to-dateness rates on Google

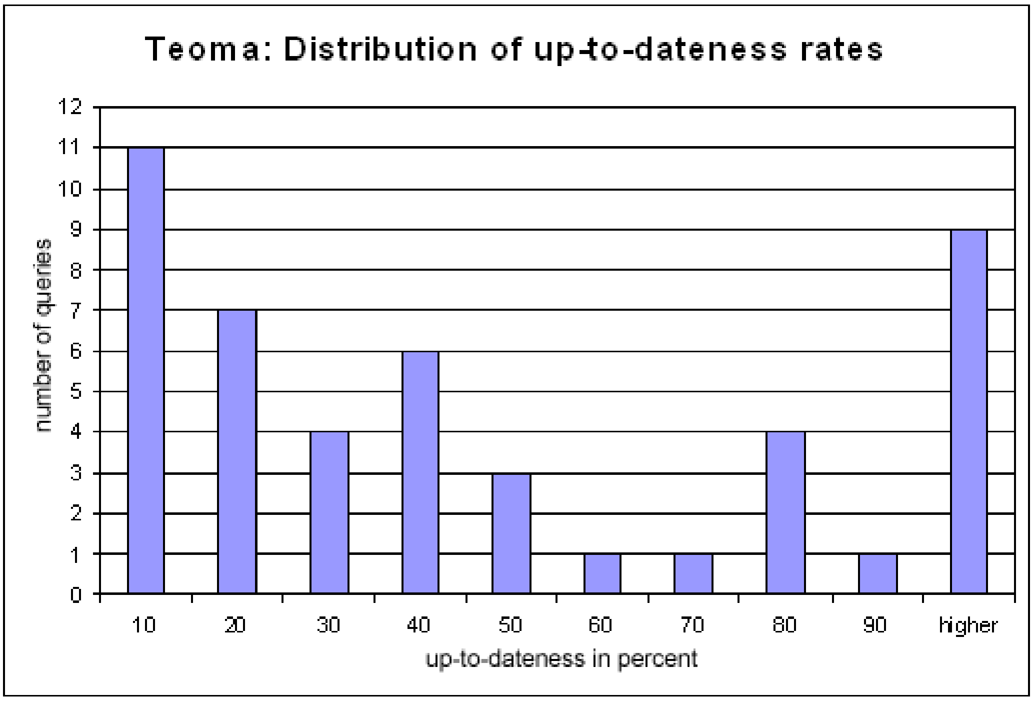

Figure 2: Distribution of up-to-dateness rates on Teoma

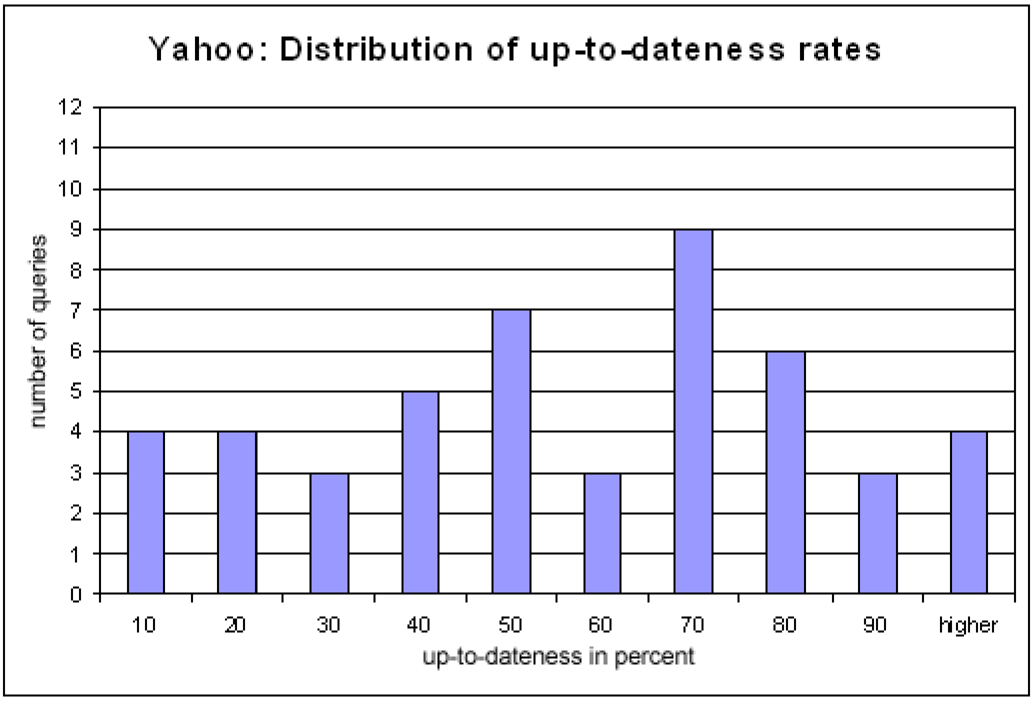

Figure 3: Distribution of up-to-dateness rates on Yahoo!Error ratio

The user is not only interested in the ratio of documents classified correctly but also in the error ratio. So far, we have only looked at the up-to-dateness rate as an indicator for the performance of the search engines. As a counterpart to the up-to-dateness rate we introduce the error ratio, which measures the ratio of documents classified incorrectly.

The error ratios of the search engines are shown in table 3. We see that Teoma classifies more documents incorrectly than correctly. The error ratio for this engine is 62.6 percent. Yahoo! performs better but has an error ratio of 45.7 percent, while Google shows an error ratio of 40.5 percent, which does not reflect a good performance either. Four out of ten documents are classified incorrectly. This means that the user obtains search engine results pages that are crammed with "noise". He might get the original results improved, but the search engines fail to provide a clean list of relevant documents. A statistical test proves the differences between the engines to be significant.

Table3: Error ratio in the analysed search engines

Search engine	Number of documents classified correctly	Number of documents classified incorrectly	Error ratio (in percent)
Teoma	115	193	62.66
Google	166	113	40.50
Yahoo!	154	130	45.77

The high error ratios of all tested search engines confirm our assumption that search engines in general have problems to determine the actual update of a document.

Seeing these dissatisfying results, we ask whether a user should use date restricted searching at all. Table 4 shows for our test queries in how many cases it would be useful to perform a date restricted search, in how many cases this would be useless and in how many cases it does not change the results. For this analysis we omitted queries that resulted in a 100 percent up-to-dateness rate for the simple as well as for the date restricted search. For example, a user interested in the Michael Jackson case might try a regular search for the topic and when he does not find recent documents he might choose to restrict his search to documents updated within the last six months. We ask whether it is useful to restrict the search.

The winner in this evaluation is Yahoo!. But here as well, only about two thirds of the queries produce better results by using the date restriction. It is interesting to see that with all search engines a relatively large fraction of date restricted queries produces poorer or unchanged results.

Table 5: Improvements and deteriorations caused by the use of date restricted queries

Search engine	deteriorations	no alteration	improvements
Teoma	14	17	16
Google	8	12	25
Yahoo!	7	10	30

Winners in individual queries

Figure 4 shows how many queries were answered in the most satisfactory manner by using a particular search engine independent of the achieved up-to-dateness rate. The best search engine in this respect is the one that gets the highest up-to-dateness rate for a specific query. We ranked the results for each query; when two search engines achieved the same rate of up-to-dateness they got the same rank. In these cases the third search engine achieved the next lower rank. When a search engine achieved an up-to-dateness rate of zero percent it was ranked third.

With 24 queries, Yahoo! is ranked first followed by Google with 18 first ranks. As mentioned earlier, Google achieved the highest up-to-dateness rate for all results but this is not the case as far as all individual queries are concerned. Looking at the winners in all individual queries we can give no recommendation for a particular search engine. Yahoo!, the winner of our test, achieves the first rank in only less than half of the queries. It seems that the criterion for a search engine to be the most efficient for the retrieval of up-to-date documents is heavily query dependent. Even Teoma, the search engine that did rather badly in our tests, came out on top in 30 percent of our queries.

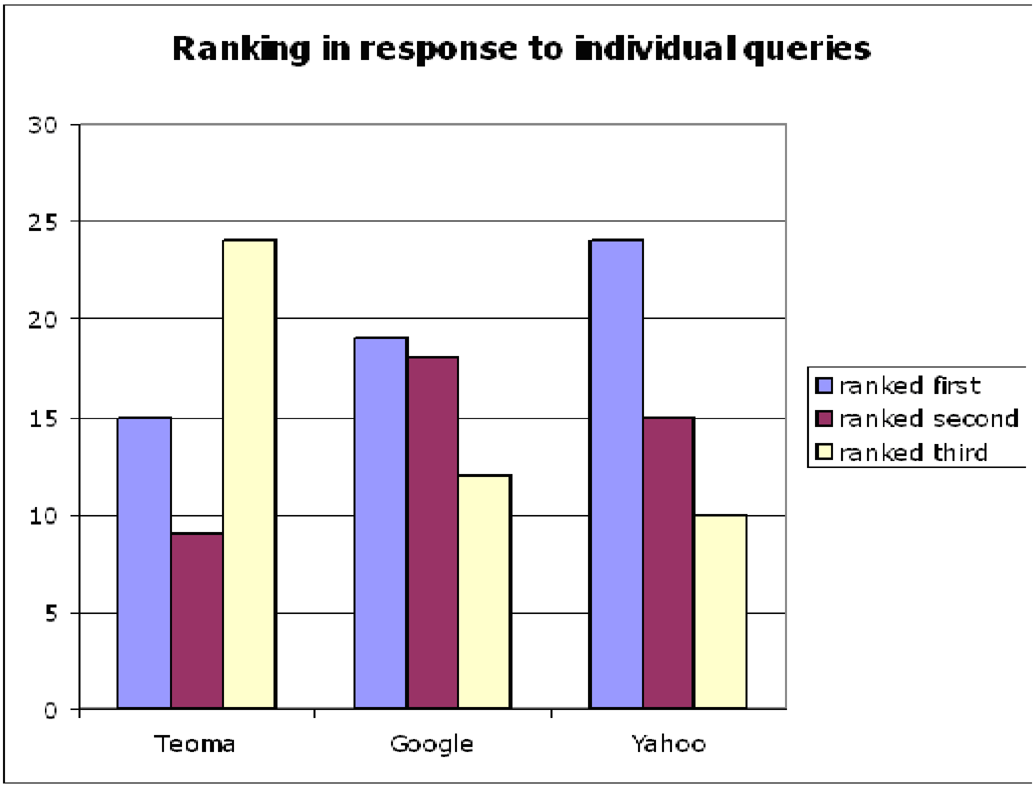

Figure 4: Ranking in response to individual queries

Discussion

The results of our study prove that date restricted searches fail to work properly in web search engines. We were able to show that Google performs best with regard to an overall up-to-dateness rate but does not perform best with each individual query.

We cannot give a clear recommendation for using one particular search engine to make date restricted queries. But that much is certain: the date restriction in Teoma should not be used. Our study revealed that Teoma is unable to improve the up-to-dateness rate significantly when using a date restriction.

It seems that every tested search engine relies on the date provided by the server and / or on the changes in the document compared to a version from a previous crawl. Therefore, they regard minor layout modifications or even a change of adverts as a change in content. To solve the problems emerging from this, search engines should consider using the degree of change of web pages in addition to the sole frequency of change. As Ntoulas, Cho and Olsten (2004) pointed out, the frequency of change is not a good indicator of the degree of change. In many cases, when web pages change, they only change in their markup or in trivial ways (Fetterly et al. 2004, 234). Today this leads the search engines to assess these changes as an actual update of the document.

Our study shows that the problems with date searching that we discussed are far from being solved. Although the problem is known in research, there is no literature discussing possible solutions to it. Works that deal with the future of web information retrieval simply suppress the problem (e.g. Chakrabarti, 2003; Henzinger, Motwani, Silverstein, 2002).

It seems obvious that it is impossible to determine the actual update of a web page just by one of the factors mentioned. The only solution to this problem seems to be a combination of the four following factors: server date, date of the document's first being indexed, meta data (where available) and update information provided in the contents of the page. Combining these factors can help fix an at least approximate date of the actual update of the document. In addition to the techniques used so far we recommend especially the extraction of date information from the contents of web pages. However, this should be used as well in combination only. We assume that most web page authors give no thought to this information being used to determine the document's update status. Therefore, the retrieved information is not consistent and should be regarded as not always trustworthy.

References

Chakrabarti, S. (2003), Mining the web: Discovering knowledge from hypertext data, Morgan Kaufmann, Amsterdam.

Fetterly, D.; Manasse, M.; Najork, M.; Wiener, J. L. (2004): A Large-Scale Study of the Evolution of Web Pages. Software – Practice and Experience 34(2), 213-237

Henzinger, M., Motwani, R., Silverstein, C. (2002), „ Challenges in Web Search Engines“, SIGIR Forum, Vol 36 No 2. Available
http://www.acm.org/sigs/sigir/forum/F2002/henzinger.pdf

Lewandowski, D. (2004), „Abfragesprachen und erweiterte Funktionen von WWW-Suchmaschinen“, IWP - Information: Wissenschaft und Praxis, Vol 55 No 2, pp. 97-102.

Machill, M.; Neuberger, C.; Schweiger, W.; Wirth, W. (2003), „Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen“, in Machill, M.; Welp, C. (Eds.), Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen, Verlag Bertelsmann-Stiftung, Gütersloh, pp. 13-490.

Ntoulas, A.; Cho, J.; Olston, C. (2004), “What’s New on the Web? The Evolution of the Web from a Search Engine Perspective”, Proceedings of the Thirteenth WWW Conference, New York, USA. Available http://oak.cs.ucla.edu/~ntoulas/pubs/ntoulas_new.pdf

Price, G.; Tyburski, G. (2002), “It's Tough to Get a Good Date with a Search Engine”, Search Day 5.6.2002. Available http://www.searchenginewatch.com/searchday/article.php/2160061

Spink, A. (2003), “Web Search: Emerging Patterns”, Library Trends, Vol 52 No 2, pp. 299-306.

Spink, A.; Jansen, B. J.; Wolfram, D.; Saracevic, T. (2002): From E-Sex to E-Commerce: Web Search Changes. IEEE Computer 35(3), 107-109

Sullivan, D. (2003), “Search Engine Sizes”. Available: http://searchenginewatch.com/reports/article.php/2156481