

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

Dirk Lewandowski¹

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 derivatives 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

Dirk Lewandowski

- 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.

Dirk Lewandowski

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

