How Google Web Search copes with very similar documents

Wouter Mettrop[1], Paul Nieuwenhuysen[a] and Hanneke Smulders[b]
[a] *Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussel, and Universiteit Antwerpen, Belgium*
[b] *Infomare consultancy, P.O. Box 333, NL-4530 AH Terneuzen, The Netherlands*

A significant portion of the computer files that carry documents, multimedia, programs etc. on the Web are identical or very similar to other files on the Web. How do search engines cope with this? Do they perform some kind of "deduplication"? How should users take into account that web search results are influenced by "deduplication"? We have investigated this deduplication function of the Google Web search engine. The focus on Google Web Search is motivated by the high popularity of this Web search engine. We developed a well-controlled experimental environment, with very similar test documents on various Web server computers in two countries and with automated scripts on a client computer. We report here the results of this investigation. We found that users may miss documents due to deduplication, and that it is not straightforward to cope with this due to complications as follows. We observed various types of deduplication and in the query result sets we noted changes/fluctuations over time. Part of these changes over time occurred only once in a series of measurements, while others were continuous, persistent, and thus more significant. This work is also motivated by the following: Variations in the contents of documents can be considered as small in deduplicating computer systems, which leads to hidden documents, while the same small variations can create quite different meanings for a human user and reader. This is probably the first investigation of deduplication in Web search from the user's point of view.

Keywords: Google Web search, very similar documents, deduplication, fluctuations

1 INTRODUCTION

How do Web search engines cope with duplicate files on the WWW? More specifically, how do identical or very similar documents show up in the results presented to users? Are documents hidden due to deduplication? If yes, then how should a user take this into account?

This is a relevant research question because

- WWW search engines have become quite important information systems with a huge user community
- a large part of files on the WWW are very similar (Fetterly et al. 2003)
- clarification is desired not only for expert users of search engines in bibliometrics /informetrics, but also for all information searchers because documents that are formally similar for automated computer systems can carry significant differences in meaning for a human reader.

For this investigation we define deduplication of search results as the selection of one or more documents to represent a cluster of all the documents that are considered as very similar by the search system. The default search result lists present what we call "representatives". Deduplication can be applied during the harvesting and processing of new Web pages or during the retrieval phase.

The length of this document is limited; therefore we focus here on our investigation of Google Web Search, because this is nowadays the most important search engine in many ways.

2 METHODS

The concept of deduplication is a relative concept. From the user's point of view, deduplication can be observed and investigated only when result sets of queries miss documents that are present ("known" by the search engine) in the result sets of other queries that are "comparable". By "comparable" queries we mean queries that search for the same documents, while using different user interfaces that are offered by the same search engine. Google offers several interfaces, which allow such a comparison. In the simple (not the advanced) search mode, Google Web Search offers two interfaces for searching on the content of Web documents: besides the default search mode there is also a search mode "with the omitted results included", which is offered via a link on the last page with search results. In the default mode, Google Web Search deduplicates the search result by omitting "very similar" entries. In the second mode, Google Web Search includes these omitted entries and presents non-deduplicated search results. Moreover Google Web Search offers an interface for searching on non-content aspects of the documents, such as language, file format, date

[1] Corresponding Author: Wouter Mettrop. CWI (Centrum voor Wiskunde en Informatica), Kruislaan 413, NL-1098 SJ Amsterdam, The Neherlands. Email wouter.mettrop@cwi.nl

and URL.

We have put 18 test documents on 8 server computers in 2 countries. It has been our intention that Google is able to recognize this set of Web documents as near-duplicates. For our investigation we worked with a set of test documents that contain no formatting differences and small differences in the content. These documents were put on the Web at the end of 2002 (8 documents) and at the end of 2003 (10 documents). Our test documents did not change over time.

Series of searches were performed in an automated procedure. The scripts of the automated queries used the non-graphical browser Lynx on a UNIX computer system. The test documents were searched during the period March–June 2004. A set of 54 different queries was submitted simultaneously (i.e. one immediately after the other in a computerised batch process), every 30 minutes:

- A subset of 18 content queries searched for words in the body text of the test document in the default way.
- Another subset was almost identical but asking to show also the search results that are omitted (hidden) by default due to the deduplication function.
- The third and last subset of 18 queries searched for words in the URL of each test document.

The search results are filtered and only the test documents are taken into account. The set of 54 queries was submitted 4993 times; so the total number of queries submitted is 269622.

All queries and results are described in http://www.cwi.nl /archive /projects /IRT /doc /clustering.html

In this article, the following terms will be used:

- **representative document(s):** Web page(s) selected by the search engine to represent a cluster of recognized copies or very similar documents;
- **representative queries**: queries that search for content in the default search mode;
- **cluster queries**: queries that search for content in the search mode that presents the whole cluster; that is the non-deduplicated search result;
- **URL queries**: queries that search for a (part of) the URLs of the test documents;
- **a set of queries**: a set of (18) queries of one of the three types of query, mentioned above, submitted simultaneously;
- **individual representative query**: a single, individual representative query from the set of 18 representative queries;
- **individual cluster query**: a single, individual cluster query from the set of 18 cluster queries.

3 RESULTS AND DISCUSSION

3.1 Deduplication of very similar documents

In order to compare the search results of two queries

1. we calculated how often the results of one of the two queries (named A) is a subset of the results of the other query (named B), and
2. we determined the extent of deduplication by calculating the percentage of documents that is absent in the deduplicated search result set.

Table 1 summarizes the data we collected.

Deduplication is measured in 99%-100% of comparisons of all types. Deduplication in the retrieval phase, measured in our experiment, is strong, in the sense that ±90% of documents are missing, when result sets of individual representative queries and sets of representative queries are compared to result sets of individual cluster queries, sets of cluster queries and sets of URL queries. In the default search mode, Google Web Search found in the experiment described above, in 99.9% of the submissions of the individual representative queries just 1 test document.

Moreover, some deduplication, is also observed in the sense that even the cluster queries miss ±8% of the documents, in comparison to URL queries. So some files that are known to the search engine do not show up in the search results, not only in the default search mode with deduplication, but even when we ask to show files that are by default omitted due to deduplication. Perhaps this is due to deduplication that occurs not in the retrieval phase but already in the harvesting and indexing phase.

Table 1: Data obtained in this investigation on deduplication by Google Web Search

Different types of queries that were compared, A versus B (in brackets: number of comparisons)	$A \subset B$ & $A \neq B$ number and percentage of comparisons average percentage of test documents in B missing in A per comparison for whole test period	$A \equiv B$ number and percentage of comparisons	$B \neq A \cup B$ number and percentage of comparisons
individual representative queries versus individual cluster queries (89874)	89698 = 99.80% 90.78%	75 = 0.083%	101 = 0.11%
individual representative queries versus sets of cluster queries (89874)	89799 = 99.92% 90.81%	75= 0.08%	0
sets of representative queries versus sets of cluster queries (4993)	4993=100% 89.32%	0	0
individual representative queries versus sets of URL queries (89874)	89667 = 99.77% 91.55%	75 = 0.083%	132 = 0.15%
sets of representative queries versus sets of URL queries (4993)	4981 = 99.76% 90.12%	0	12 = 0.24%
individual cluster queries versus sets of URL queries (89874)	88885 = 98.90% 8.14%	72 = 0.08%	917 = 1.02%
sets of cluster queries versus sets of URL queries (4993)	4940 = 98.94% 8.02%	0	53 = 1.06%

3.2 Web search results in the case of very similar documents change over time

The document shown after deduplication in the default search mode was not always the same document. We found that during this experiment (a couple of months) Google Web Search changed the representative document between 12 and 28 times per query.

Users who find a document in a result set may expect to find this document again when the query is repeated. However documents come and go. Deduplication itself is not the only phenomenon that makes users miss expected documents in result sets.

In an earlier investigation we described this phenomenon as “fluctuations” in search results, and we proved that expected documents/files are not always present (or counted) in the result sets of popular Web search engines, due to these fluctuations (Mettrop and Nieuwenhuysen, 2001). “Document fluctuations” show the phenomenon that a set of queries, submitted regularly, stops retrieving one or more documents that still exist in reality. Each result set is compared to a frame of reference that is made up by the preceding observation combined with knowledge about the WWW. So, analogous to the concept of deduplication, the concept of fluctuations is a relative concept. We define queries to be comparable when they search through time for the same documents using the same search interface offered by the search engine. In our investigations this interface was the interface either for representative queries, for cluster queries or for URL queries.

All experimental results are shown in table 2.

Our set of test documents and our queries do not change over time; nevertheless, our experiment reveals changes in the result sets: document fluctuations appeared in both duplicated and non-deduplicated result sets. Most document fluctuations appear in the query result sets of the (sets of) the (deduplicated) representative queries, but most documents are missing due to document fluctuations in the non-deduplicated results sets, i.e. in the result sets of (sets of) cluster queries and URL queries.

Table 2 Document fluctuations observed in the experiment with Google Web Search

Observation type (In brackets: number of observations)	Number and percentage of observations influenced by document fluctuations		Number and percentage of query set submissions, which include observations that show document fluctuations		Average number and percentage of test documents missing per			
					fluctuation		observation	
individual representative queries (89874)	329	0.37%	119	2.38%	1	98.78%	0.004	0.362%
individual cluster queries (89874)	187	0.21%	109	2.18%	8.59	64,24%	0.018	0.134%
sets of representative queries (4993)	25	0.5%			1.08	54.68%	0.005	0.274%
sets of cluster queries (4993)	3	0.06%			1	0.09%	0.001	0.0001%
sets of URL queries (4993)	74	1.52%			1.59	11.95%	0.0242	0.177%

3.3 Persistency and impact of changes over time

Are the fluctuations reported above important? Document fluctuations occur with all the types of queries that we used. In the case of deduplicated result sets, they correspond with changes of the representative documents.

Figure 1 shows the turnover of the representative documents in the results of the individual representative queries between day 50 and day 75 of the experiment.

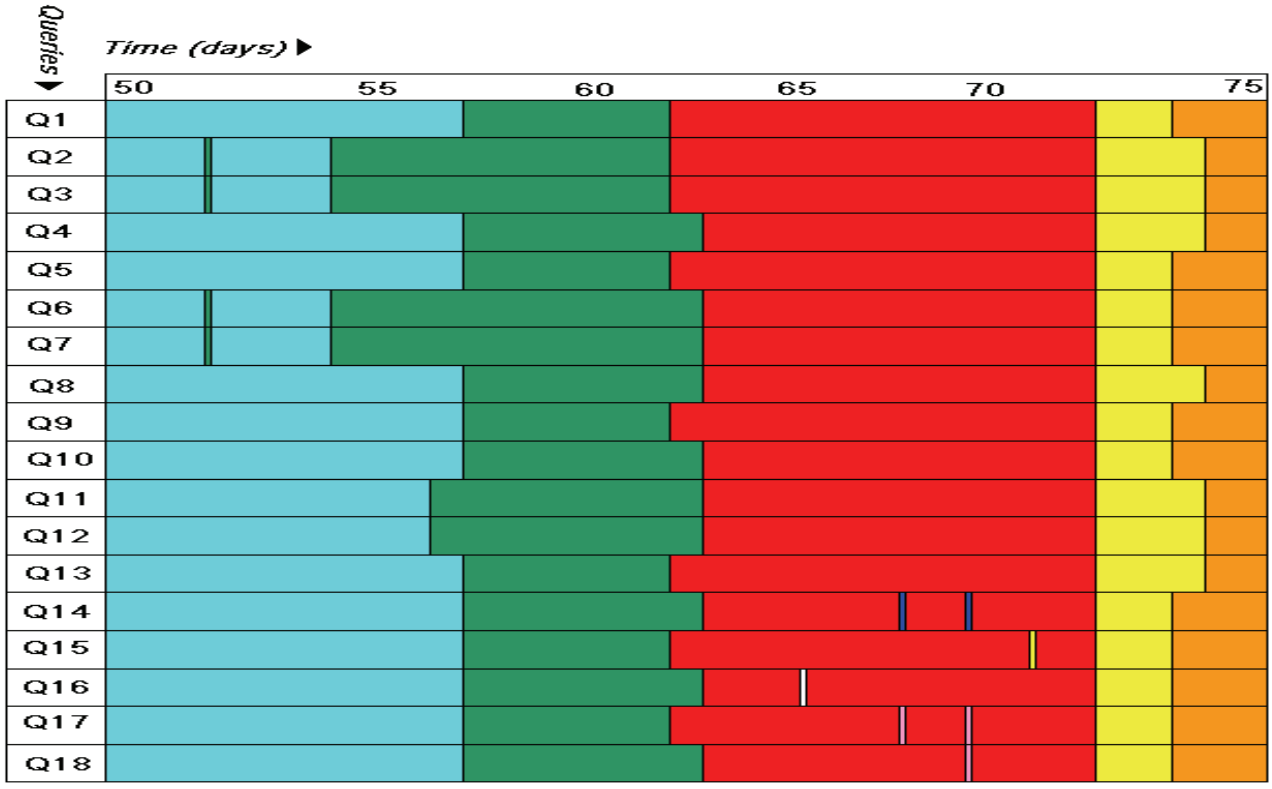

Figure 1: The particular representative test documents that were found by all 18 individual representative queries between day 50 and 75 (1200 observations). Each shade indicates another document. The thin lines represent single observations with deviating results and the thin white line indicates a result set without test documents.

All result sets contain just 1 representative test document. During the whole experiment, for each query, the representative document changed from 12 up to 28 times. In our investigation with individual

representative queries, 9 different representative documents show up 438 different periods.

We distinguish changes of the representative documents and other document fluctuations. Changes of the representative documents show reality, i.e. the results as were meant by the search engine, while the other fluctuations show the deviations in our observations of reality: the discontinuous, erratic samples caused by disorders and imperfections in the communication between search engine and end user. Duration should, among other things, reflect/determine if a change of representative document is only an erratic document fluctuation or part of the more continuous reality. In the case of individual representative queries, 107 document fluctuations (changes in the result set) lasted only one observation. These include 68 observations that found nothing. The other 39 observations found a deviating test document. Other periods lasted between 2 observations, i.e. 1.5 hour at the most (9 times) and 1296 observations, i.e. 27 days (also 9 times). It never occurred that a query found nothing for 2 succeeding observations.

Another aspect is the number of queries within the same observation, submitted at the same time, that retrieve the same document. All 39 queries with observations mentioned above, which found a deviating test document that was found only by a minority of queries within the same observation, and these queries also found the test document only during one observation.

In the case of individual representative queries in the experiment it seems reasonable to consider changes of the representative document that persisted at least 2 observations as real changes of the representative document, and changes that lasted only one observation as other erratic fluctuations. In this view, changes of the representative documents occurred 178 times and other document fluctuations occurred 151 times. Changes of representative documents occurred 10 a 12 times for each query.

For each type of query it is possible to model the search results and to draw the distinction between changes of the showed documents and other document fluctuations. Facing the obligatory limitation in the length of this paper, we restrict this report to the modeling of individual representative queries. Moreover, we do not cover the paradoxical variations in search results, due to the fact that not all queries change representative documents the same way, as can be seen in figure 1, which we named element fluctuations in Mettrop and Nieuwenhuysen (2001).

4 CONCLUSIONS

Google deduplicated Web search results in various ways. Results of queries submitted in the default search mode were deduplicated very well. They show only 1 test document, which is ±10% of the test documents that are shown with the simultaneous submitted analogous query that asks for the omitted search results also. But Google also deduplicates the results of queries that ask for the omitted search results, as more test documents were found while simultaneously searching for the terms in the URL.

Moreover, deduplicated and non-deduplicated result sets show document fluctuations; most of these occur in the result sets of sets of URL queries. Some of them are due to the search engine that changes the presented documents, and some of them are due to erratic imperfections in the communication between end user and search engine. The number of document fluctuations is small, but users should be aware of such changes over time

REFERENCES

[1] Fetterly, D. M. Manasse and M. Najork. On the evolution of clusters of near-duplicate web pages. In Proceedings of the First Latin American web Congress. 2003. Published by IEEE nr 0-7695-2058-8/03.

[2] Mettrop, W., and Nieuwenhuysen, P., Internet search engines – Fluctuations in document accessibility. *J. Documentation*, 57 (5):623-651, 2001.