THE RELIABILITY OF INTERNET SEARCH ENGINES: FLUCTUATIONS IN DOCUMENT ACCESSIBILITY

Wouter Mettrop, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, The Netherlands and Paul Nieuwenhuysen, Vrije Universiteit Brussel and Universitaire Instelling Antwerpen, Belgium

Keywords: Internet, WWW, Web, Search Engines, Information Retrieval, Performance Tests, Depth Of Indexing, Fluctuations, Experimental Investigations

Abstract: An investigation of the consistency of retrieval through Internet search engines is reported. Search engines are often compared on the basis of their sizes; i.e. the number of documents indexed in their databases. But, does a bigger size always result in more hits? This presentation shows an empirical evaluation of 13 engines: AltaVista, EuroFerret, Excite, HotBot, InfoSeek, Lycos, MSN, NorthernLight, Snap, WebCrawler and three national Dutch engines: Ilse, Search.nl and Vindex. The focus is on a characteristic related to size: the degree of consistency to which an engine retrieves documents. Does an engine always present the same relevant documents that are available in its database? Three types of possible fluctuations in the result sets of search engines are identified. In many cases, such fluctuations are significant; so they should be taken into account by users of an Internet search engine and in other research in the domain of the performance of Internet search engines.

1. INTRODUCTION

1.1. The WWW and Search Engines

The amount of information that is potentially available straightforward through the Internet, keeps on growing. Estimates lead to the conclusion that now, in the year 2000, about 1 billion, that is 1 000 million, unique URLs or 'pages' are accessible in the total Internet (see for instance http://searchengineshowdown.com/, Lawrence and Giles, 1999, and http://www.inktomi.com/webmap/ cited February 2000). Those documents correspond to about 10 terabyte (about 10 000 gigabyte) of text data. Clearly, some good retrieval systems are required to enhance the value of this unordered collection of information resources. Many systems are already available. Simplifying reality somewhat, we can distinguish several types:

- Directories of selected sources categorised by subject, made by humans, mainly for browsing.

- Search engines, based on databases with machine made indexes, mainly for word-based searching.
- "Meta-search" or "multi-threaded" search systems that rely mainly on the databases made available by the previous systems, for word-based searching.

During the last few years, an investigation of the performance of several well-known international (and a few smaller local) word-based Internet search engines has been going on, by a group of information professionals from various institutes and companies in The Netherlands (and Belgium), as named in the Appendix. Here we report on a part of this work.

Many aspects/criteria can be considered in the evaluation of an Internet search engine, including
- the coverage of documents present on the WWW (and most of the well-known investigations have focused on this aspect already, including Lawrence and Giles, 1999)
- the elements of a document, that are indexed to make them usable for retrieval,
- the absence of dead links in the set of links suggested by the search system,
- fluctuations and inconsistencies in the result sets offered by a search engine, that may exist, and that complicate the retrieval proces.

We started by studying the depth of indexing of some search engines and in this way were confronted with fluctuations that exist in the performance of most systems. This phenomenon is relatively unknown, obscure and not well investigated, but it may hinder and influence serious quantitative investigations of other aspects. Therefore we have made what we think is the first quantitative study in this area. This work has pointed out that most engines suffer from this 'incorrect variable behaviour', in the sense that unexpected and annoying fluctuations exist in the result sets of documents, which means that documents are not retrieved reliably. The first phase of our investigation has been reported earlier (Mettrop and Nieuwenhuysen, 1999).

1.2. Fluctuations in the Performance of Search Engines

The result set of documents, shown by an Internet search engine as response to a query, changes over time. Broadly speaking, an alteration in this set is correct, if it is a reflection of an alteration in the WWW reality, as documents are added to, removed from or changed in the WWW. If not, the change can be seen as incorrect. Such incorrect changes do not only concern incorrect removals of documents from the set of indexed documents, or incorrect (late) additions to this set. They can also occur when an engine has indexed a document, but not always succeeds afterwards in retrieving it.

A fluctuation appears when the result set of an observation (i.e. one query or a set of queries) does not come up to expectations. What is expected depends on other information, obtained from other observations and from knowledge about the reality of the WWW. For example, when an engine did retrieve a certain document on the base of a certain query, and when you know that this document still exists, then you expect to find this document again when you submit the query again.

A fluctuation can be an increase in the number of documents in the result set, or a decrease.

In this investigation we only consider decreases. We investigate the disappearance of documents from result sets,not the addition of documents to result sets, because adding documents to a search engine's database is a natural and normal process for an Internet search system.

Definition: A fluctuation appears, when the result set of an observation (i.e. one query or a set of queries) misses documents with respect to a certain expected result set.
Here, one 'observation' is one query or a set of queries and the expected result set is determined by a certain frame of reference: observations and/or knowledge about web reality.

1.3. How to Detect Fluctuations?

We consider and apply two methods in comparing result sets:

- Comparison through time: Compare the result sets of one observation that is repeated regularly with an interval of a few days.
- Comparison at one moment: Compare the result sets of all queries in a set of queries that are submitted at the same time.

1.4. Types of Fluctuations

According to the different methods for detecting fluctuations, we distinguish three types of fluctuations, as follows:

Document fluctuations:

- Comparison through time.
- One observation means here a set of queries.
- Frame of reference for each observation: other observations and knowledge of WWW.

A fluctuation appears when one of the total result sets, belonging to one observation, does not mention a document that was mentioned previously and that did not disappear in reality from the WWW.
We call these fluctuations 'document fluctuations', because it seems as if search engines 'forget' whole documents.

Element fluctuations:

- Comparison at one moment.
- One observation means here one query in a set of queries.
- Frame of reference for each observation: other queries in the set.

A fluctuation appears, when the results set of one of the queries mentions some, but not all, identical documents that are retrieved by all queries in the set. Here we need a set of identical documents to search for. We do not see a simple explanation for this type of fluctuation.
We call a fluctuation of this type an 'element fluctuation', because it corresponds with a certain query, which corresponds with a certain document element.
While detecting this type of fluctuation is only possible in an artificial environment with identical documents, the phenomenon of course does occur in all practical situations.

Indexing fluctuations:

- Comparison through time.
- One observation means here one query.

- Frame of reference for one observation: other observations and knowledge of WWW.

In this case, single queries are compared through time. Because these observations can be influenced by element fluctuations defined above, they are not useful for detecting document fluctuations. What can be observed, is the indexing policy of an engine with respect to the query. By analyzing all the result sets of one query, it is possible to establish if an engine indexes the document-element corresponding with the query or not, and if there are fluctuations in this indexing policy. We then call these 'indexing fluctuations'.

2. METHODS AND PROCEDURES OF OUR TEST

2.1. Search Engines Investigated

13 Internet search engines have been investigated: AltaVista, EuroFerret, Excite, HotBot, InfoSeek, Lycos, MSN, NorthernLight, Snap, WebCrawler and three Dutch engines: Ilse, Search.nl and Vindex. Some of the investigated engines have several search modes: simple, advanced, super, fuzzy etc. We always used the most simple mode and we did not change the mode during the tests.

2.2. Test Documents

16 identical test documents were placed on different sites in The Netherlands and in Belgium; 7 documents were submitted to the engines and linked to a document known by the engines. These 16 documents remained unchanged during the testing period.

2.3. Test Queries

32 queries were formulated so that each one is related to a specific element of the test document. These were repeatedly and regularly submitted: 1 per 29 minutes. One round of 32x13 queries (all queries to all engines) took 9 days. 43 rounds were included in the experiment reported here, during 14 months, from October 1998 up to December 1999.

2.4. Measurements Made

Observed / counted / measured were the following:

- which of the identical documents were retrieved by each engine,
- which document elements were indexed by each engine,
- the number of document, element and indexing fluctuations per engine, and
- the size of these fluctuations.

2.5. Definition of Fluctuations in the Test

- 'Document fluctuations' appear, when engines forget test documents, i.e. a test document, found in round n, has disappeared in round n+1 (for n=1 to 42 in this investigation).
  Here one observation means one complete round of 32 queries. The frame of reference is the result of the preceding round. The size of a document fluctuation is here the number

of documents missing, expressed as percentage of the number of documents expected; so: 0 < size =< 100.

- 'Element fluctuations' appear when the 32 test queries, submitted to an engine in one round, do not lead to result sets containing the same (non empty) subset of test documents.
  An observation is here one specific query in one specific round. The frame of reference is the result of all 32 queries within this round. The size of an element fluctuation: the number of documents missing, expressed as percentage of the number of documents expected; so: 0 < size =< 100. Notice: According to the definition, an element fluctuation should be detected in a comparison at one moment. Here, for technical reasons, we did not submit all 32 queries at one moment, but one after the other in 32x29 minutes = 15 hours.
- An 'indexing fluctuation' appears when query m in round n finds no documents, while the same query m found a document in round n-1, and will not find documents in following rounds n+1, n+2 and n+3. An observation means in this case one specific query in one specific round. The frame of reference is here the same query, in the preceeding round and in the following 3 rounds. The size of each indexing fluctuation is 100, as all expected documents are missing.

Mutual influences of fluctuations have been excluded.

3. RESULTS OF OUR TEST

Search engine	Number of test documents linked / submitted	Number of test documents not linked / submitted
WebCrawler	02	00
Vindex	07	01
Snap	07	02
Search.nl	05	00
NorthernLight	03	00
MSN	03	01
Lycos	03	00
InfoSeek	02	00
Ilse	05	00
HotBot	07	02
Excite	06	00
EuroFerret	05	01
AltaVista	04	01

Table 1: Number of the identical test documents that were retrieved during the whole test period

Search engine	Percentage of result sets that missed documents (lower is better)	Average percentage of missed documents per result set (lower is better)
WebCrawler	07%	04%
Vindex	93%	23%
Snap	10%	05%
Search.nl	01%	01%
NorthernLight	00%	00%
MSN	57%	30%
Lycos	00%	00%
InfoSeek	05%	03%
Ilse	13%	09%
HotBot	56%	23%
Excite	07%	04%
EuroFerret	00%	00%
AltaVista	07%	05%

Table 2: Element fluctuations

Search engine	Percentage of rounds that missed documents (lower is better)	Average percentage of missed documents per round (lower is better)
WebCrawler	20%	17%
Vindex	21%	03%
Snap	43%	14%
Search.nl	00%	00%
NorthernLight	06%	06%
MSN	16%	12%
Lycos	07%	05%
InfoSeek	03%	03%
Ilse	00%	00%
HotBot	26%	09%
Excite	35%	34%
EuroFerret	19%	07%
AltaVista	07%	05%

Table 3: Document fluctuations

Search engine	Percentage of result sets that missed documents = Average percentage of missed documents per result set (lower is better)	Number of indexed elements; measured over the complete test period
WebCrawler	0.0%	19
Vindex	1.2%	28
Snap	0.6%	18
Search.nl	0.0%	18
NorthernLight	0.0%	16
MSN	1.7%	27
Lycos	0.0%	17
InfoSeek	1.5%	13
Ilse	0.0%	16
HotBot	0.7%	18
Excite	0.0%	16
EuroFerret	0.0%	04
AltaVista	0.4%	20

Table 4: Indexing fluctuations

Search engine	Document fluctuations	Element fluctuations	Indexing fluctuations
WebCrawler	17%	04%	00.0%
Vindex	03%	23%	01.2%
Snap	14%	05%	00.6%
Search.nl	00%	01%	00.0%
NorthernLight	06%	00%	00.0%
MSN	12%	30%	01.7%
Lycos	05%	00%	00.0%
InfoSeek	03%	03%	01.5%
Ilse	00%	09%	00.0%
HotBot	09%	23%	00.7%
Excite	34%	04%	00.0%
EuroFerret	07%	00%	00.0%
AltaVista	05%	05%	00.4%

Table 5: Average percentage of test documents missed per observation due to fluctuations

4. DISCUSSION AND CONCLUSIONS

4.1. Search Engines Make Mistakes

The Internet retrieval systems investigated

- forget documents completely (document fluctuations),
- miss documents in their result sets (element fluctuations), and
- are subject to changes in indexing policy (indexing fluctuations).

Is this important?
When searching for a known item, one should be aware of this (especially users who try to find back a document). In the case of an engine with element fluctuations one should repeat the search with other search terms and find that result sets can be inconsistent. In the case of an engine with document fluctuations, one should repeat the same query later.
The percentage of missed documents, together with other reducing effects, determines the effective size / coverage of an engine.
The influence of indexing fluctuations can be neglected in comparison with the influence of element and document fluctuations. The influence of document fluctuations is comparable to the influence of element fluctuations.

4.2. What is Correct and What is Incorrect?

The fluctuations have one thing in common: users may miss documents, including documents that they expect to retrieve. The influence of each type of fluctuation is different (which is measured in this investigation) and so is the degree to which search engines should be blamed for it. In the introduction above, mistakes are mentioned: Is every fluctuation we define a mistake?
One can say that fluctuations are correct if they are a reflection of alterations in the WWW reality (documents are added or removed). If not, they are incorrect. With this condition, all of the three types of fluctuations that we consider are incorrect.
Another definition could say that fluctuations are correct if they are a reflection of alterations in the indexed database of an engine. In that case, document fluctuations and indexing fluctuations are correct, but element fluctuations are still incorrect.
In practice most users will not care about what is 'correct' or 'incorrect'; they simply cannot find a document that still exists in reality. If it is not their own mistake, it is the engine's mistake.

4.3. Comparison of Search Engines, Concerning Fluctuations

Document fluctuations have been found in the result sets of all search engines except for two of the search engines restricted to The Netherlands: Ilse and Search.nl. Most document fluctuations were shown by Snap and Excite. Excite also misses most documents per round by document fluctuations.
Element fluctuations have been found in the result sets of all search engines except EuroFerret, Lycos, NorthernLight and (almost) Search.nl. Most element fluctuations were shown and most documents per result set were missed by element fluctuations by Vindex, MSN and HotBot.

Some engines were hardly or not subject to document or element fluctuations: Search.nl., NorthernLight, Lycos, Infoseek, AltaVista. Their result sets were complete and did not change. Moreover, Lycos, NorthernLight and Search.nl did not change their indexing policy during our research.

Many indexing, document, and element fluctuations were observed with Vindex, MSN and HotBot. Moreover, they missed the highest percentage of documents by element fluctuations. Excite forgets the highest percentage of documents by document fluctuations.

4.4. Relations Concerning Fluctuations

- Search engines that showed many element fluctuations showed also many document fluctuations. Moreover, they index many document elements and showed many indexing fluctuations.
- Engines showing many document fluctuations did not always show many element fluctuations.
- Engines that index few document elements showed few element fluctuations.

5. REFERENCES

1. Lawrence, Steve, and Giles, C. Lee. "Accessibility of information on the web." *Nature*, Vol. 400, 107-109, 8 July 1999.

2. Mettrop, W., and Nieuwenhuysen, P. "Some empirical research on the performance of Internet search engines". *In: Digital Libraries: Interdisciplinary Concepts, Challenges and Opportunities. Proceedings of CoLIS 3, the Third International Conference in Library and Information Science, Inter-University Centre Dubrovnik*, edited by: Tatjana Aparac, Tefko Saracevic, Peter Ingwersen and Pertti Vakkari, Zagreb, Lokve, Croatia : Benja Publishing, 1999, pp. 329-333.

6. APPENDIX: ABOUT THE AUTHORS

- Wouter Mettrop - CWI, Kruislaan 413, 1098 SJ Amsterdam, the Netherlands Wouter.Mettrop@cwi.nl
- Paul Nieuwenhuysen, Vrije Universiteit Brussel, Pleinlaan 2, B-1050, Brussel, Belgium pnieuwen@vub.ac.be

The following persons have been involved in some way and in some phase of the research project on Internet information retrieval tools (in alphabetical order):

- Louise Beijer (Hogeschool van Amsterdam, The Netherlands)
- Hans de Bruin (Unilever Research Laboratorium, Vlaardingen, The Netherlands)
- Hans de Man (JdM Documentaire Informatie, Vlaardingen, The Netherlands)
- Rudy Dokter (PNO Consultants, Hengelo, The Netherlands)
- Marten Hofstede (Rijksuniversiteit Leiden, The Netherlands)
- Wouter Mettrop (CWI, Amsterdam, The Netherlands)

- Paul Nieuwenhuysen (Vrije Universiteit Brussel and Universitaire Instelling Antwerpen, Belgium)
- Eric Sieverts (Hogeschool van Amsterdam, and Rijksuniversiteit Utrecht, The Netherlands)
- Hanneke Smulders (Infomare, Terneuzen, The Netherlands)
- Hans van der Laan (Computer and Internet consultant, Leiderdorp, The Netherlands)
- Ditmer Weertman (ADLIB, Utrecht, The Netherlands)